Міністерство освіти і науки України
Національний університет "Запорізька політехніка"

Субботін С.О., **Олійник А.О.**,
Льовкін В.М., **Леощенко С.Д.**

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ, ФРЕЙМВОРКИ ТА ПРОГРАМНІ ЗАСОБИ ДЛЯ ПРОГНОЗУВАННЯ І ДІАГНОСТУВАННЯ НЕЛІНІЙНИХ ОБ'ЄКТІВ

Монографія

Запоріжжя
2023

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Національний університет “Запорізька політехніка”

Субботін С.О., *Олійник А.О.*,
Льовкін В.М., *Леощенко С.Д.*

Інтелектуальні методи, фреймворки та програмні засоби для прогнозування і діагностування нелінійних об’єктів

Монографія

Видання підготовлено в межах науково-дослідних робіт «Інтелектуальні методи та засоби діагностування та прогнозування стану складних об’єктів» (номер державної реєстрації 0122U000972) та «Розроблення методів та засобів для аналізу та прогнозування динамічної поведінки нелінійних об'єктів» (номер державної реєстрації 0121U107499)
за часткової підтримки міжнародних проєктів "EuroPIM Virtual Master School Ukraine (EU-ViMUk) Німецької служби академічних обмінів DAAD та "Cross-domain competences for healthy and safe work in the 21st century" (WORK4CE, Project Reference: 619034-EPP-1-2020-1-UA-EPPKA2-CBHE-JP) програми Еразмус+ Європейського Союзу

DAAD
Deutscher Akademischer Austausch Dienst
German Academic Exchange Service

Co-funded by the Erasmus+ Programme of the European Union

WORK4CE

Запоріжжя
2023

УДК 004.891.3:004.855.5
І 73

Рекомендовано до друку Вченою радою
Національного університету "Запорізька політехніка"
(протокол №4 від 27.11.2023 року)

Колектив авторів:
Субботін С. О., д.т.н., проф. (розділ 1)
Олійник А. О., д.т.н., проф.. (розділи 2, 3)
Льовкін В.М., к.т.н.,доцент (розділ 4)
Леощенко С.Д., доктор філософії (розділи 2, 3)

Рецензенти:
Шаховська Н.Б., завідувач кафедри систем штучного інтелекту Національного університету "Львівська політехніка", доктор технічних наук, професор;
Остапов С.Е., завідувач кафедри програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича, доктор фіз.-мат. наук, професор.

І 73 **Інтелектуальні** методи, фреймворки та програмні засоби для прогнозування і діагностування нелінійних об'єктів : монографія / [С. О. Субботін, А. О. Олійник, В. М. Льовкін, С. Д. Леощенко]. – Запоріжжя : Національний університет "Запорізька політехніка", 2023. – 186 с.

ISBN 978–617–529–432–1

Монографія містить аналіз та дослідження інтелектуальних методів, фреймворків та програмних засобів для прогнозування і діагностування нелінійних об'єктів. Наведено кластеризації даних на основі індуктивного навчання нейро-нечіткої мережі з гешуванням відстаней. Розроблено нейроеволюційний метод для організації пошуку аномалій у часових рядах. Розглянуто використання рекурентних нейронних мереж для інформаційно-орієнтовних застосунків. Розроблено фреймворки побудови програмних засобів прийняття рішень для медичного діагностування.

Видання призначено для наукових співробітників, аспірантів, студентів комп'ютерних спеціальностей закладів вищої освіти, а також може використовуватися педагогічними працівниками та практичними фахівцями.

ISBN 978–617–529–432–1

ЗМІСТ

ВСТУП

Завдання прогнозування та діагностування нелінійних об’єктів є важко формалізованими завданнями та в даний час є особливо актуальними у зв'язку з необхідністю автоматизації процесів прийняття рішень для технічних та біомедичних систем. Тому до цього часу продовжується розробка і програмна реалізація продуктивних інтелектуальних методів, фреймворків та програмних засобів для прогнозування і діагностування. Для вирішення завдань цього класу широкого розповсюдження набули методи та технології на основі штучного інтелекту та машинного навчання (нейроеволюція та штучні нейронні мережі).

У цій монографії розглядаються питання розроблення та дослідження інтелектуальних методів, фреймворків та програмних засобів для прогнозування і діагностування нелінійних об’єктів.

Перший розділ присвячено кластеризації даних на основі індуктивного навчання нейро-нечіткої мережі з гешуванням відстаней. Він містить виклад методу кластерного аналізу багатовимірних даних, який для кожного екземпляра розраховує свій геш на основі відстані до умовного центру координат, використовує одновимірну координату за віссю гешу для визначення відстаней між екземплярами, розглядає результуючий геш як псевдовихід ознаки, розбиваючи його на інтервали, з якими порівнює мітки псевдокласів-кластерів, отримуючи грубе чітке розділення простору ознак і екземплярів вибірки, автоматично генерує розбиття вхідних ознак на нечіткі терми, визначає правила віднесення екземплярів до кластерів і, як результат, формує систему нечіткого виведення класифікатора Мамдані-Заде, яка донавчається у формі нейронечіткої мережі для забезпечення прийнятного значення функціоналу якості кластеризації.

Другий розділ висвітлює нейроеволюційний метод для організації пошуку аномалій у часових рядах. Він містить дослідження та порівняльний аналіз існуючих стратегій та методів, що вирішують проблему виявлення та класифікації аномалій, а також запропоновано метод виявлення, заснований на нейроеволюційних методах. Запропоновано новий нейроеволюційний метод для організації пошуку аномалій у часових рядах. На етапі навчання метод обробляє і розділяє дані про поведінку системи. У режимі синтезу метод поступово коригує моделі, щоб в майбутньому отримати з них остаточне рішення. Отримана модель синтезується з використанням

рівномірного схрещування, що дозволяє збільшити розмір батьківського пулу з двох особин до набагато більшого числа

Третій розділ описує використання рекурентних нейронних мереж для інформаційно-орієнтовних застосунків. Питання вибору найбільш підходящої топології та типу нейронних мереж залишається надзвичайно актуальним. Найкращі результати демонструють рекурентні нейронні мережі. Цей розділ присвячено огляду та пропозиціям щодо використання рекурентних нейронних мереж і інформаційно-орієнтованих застосунках. Проведені тести та аналіз їх результатів продемонстрували актуальність використання сучасних архітектур штучних нейронних мереж.

Четвертий розділ присвячено фреймворкам побудови програмних засобів прийняття рішень для медичного діагностування. У даному розділі було запропоновано фреймворк прийняття рішень для медичного діагностування, що дозволяє об’єднати створені моделі в єдину систему, визначити взаємодію між ними та створити єдине сховище даних. Проведено експериментальне дослідження, яке в підсумку охоплювало задачі прогнозування автомобільного трафіку та прогнозування забрудненості атмосферного повітря.

Монографію підготовлено в межах науково-дослідних робіт «Інтелектуальні методи та засоби діагностування та прогнозування стану складних об’єктів» (номер державної реєстрації 0122U000972) та «Розроблення методів та засобів для аналізу та прогнозування динамічної поведінки нелінійних об'єктів» (номер державної реєстрації 0121U107499), що виконуються на кафедрі програмних засобів Національного університету “Запорізька політехніка”, за інформаційної підтримки міжнародних проєктів Німецької служби академічних обмінів DAAD "EuroPIM Virtual Master School Ukraine (EU-ViMUk) та "Cross-domain competences for healthy and safe work in the 21st century" (WORK4CE, Project Reference: 619034-EPP-1-2020-1-UA-EPPKA2-CBHE-JP) програми Еразмус+ Європейського Союзу.

Монографія містить результати оригінальних досліджень авторів і може бути рекомендована для широкого кола прикладних фахівців в області комп’ютерних наук та інформаційних технологій, технічного діагностування, розпізнавання образів. Видання орієнтоване на наукових співробітників, аспірантів, студентів комп’ютерних спеціальностей закладів вищої освіти, а також може використовуватися педагогічними працівниками та практичними фахівцями.

РОЗДІЛ 1
КЛАСТЕРИЗАЦІЯ ДАНИХ НА ОСНОВІ ІНДУКТИВНОГО НАВЧАННЯ НЕЙРО-НЕЧІТКОЇ МЕРЕЖІ З ГЕШУВАННЯМ ВІДСТАНЕЙ

Кластерний аналіз [1-17] широко використовується для аналізу даних різної природи та розмірності. Метою кластерного аналізу є розділення початкової вибірки спостережень (екземплярів) на компактно розташовані групи екземплярів або виявлення компактних областей групування екземплярів – кластерів у просторі ознак, які описують екземпляри вибірки.

За типом функцій належності, що використовуються для екземплярів до кластерів, відомі методи кластерного аналізу поділяються на чіткі [1-17] і нечіткі методи [18-26].

На відміну від чітких методів [1-17], які забезпечують більш грубе розділення екземплярів, нечіткі методи [18-26] дозволяють більш адаптивний вибір кластерів у просторі ознак, оскільки вони є високоадаптивними, але вимагають великої кількості параметрів для налаштування.

Крім того, добре відомі методи нечіткого та чіткого кластерного аналізу [1-29] характеризуються низькою швидкістю та вимогливими до ресурсів пам’яті комп’ютера через необхідність обчислення попарних відстаней між екземплярами в багатовимірному просторі ознак. Крім того, результати відомих методів кластерного аналізу [1-29] є складними для людського сприйняття та аналізу великою кількістю особливостей.

Добре відомі методи чіткого кластерного аналізу [1-17, 27-29] припускають, що початкова вибірка спостережень розділена на кластери таким чином, що кожен екземпляр належить лише одному кластеру, а розбиття формується ітераційно з початкового випадкового або визначеного користувачем розбиття до остаточного розбиття, яке задовольняє заданому критерію якості. Основними відмінностями між відомими методами чіткого кластерного аналізу [1-17, 27-29] є метод обчислення відстані, критерій якості розбиття, метод генерації початкового розбиття (набір початкових розбиттів), метод генерації нового розбиття (множини розбиттів) вибірки на

основі існуючих (або раніше розглянутих) критеріїв завершення пошуку. У цьому випадку критерій якості розбиття є функцією, яка визначається на основі попарних відстаней між екземплярами вибірки, а також відстаней від екземплярів до центрів кластерів у просторі ознак. Обчислення таких відстаней для вибірок великих розмірів є обчислювально витратним завданням, а також потребує значних витрат пам'яті для завантаження всієї вибірки спостережень у пам'ять. Крім того, завдання ускладнюється необхідністю попарного переобчислення відстаней між екземплярами.

Відомі методи нечіткого кластерного аналізу [18-26] припускають, що кожен екземпляр вибірки належить до всіх кластерів, але при різних значеннях функції належності розбиття екземплярів на кластери формується ітераційно з початкових випадкових або заданих користувачем розбиттів до остаточного розбиття, яке задовольняє заданому критерію якості. Основними відмінностями між відомими методами нечіткого кластерного аналізу, як і для методів чіткого кластерного аналізу, є метод розрахунку відстані, критерій якості розбиття, метод генерації початкового розбиття (набір початкових розбиттів), метод генерації нового розбиття (множини розбиттів) вибірки на основі існуючих (або розглянутих раніше) критеріїв для припинення пошуку. Водночас, на відміну від чіткого кластерного аналізу, нечіткі методи оперують членством (належністю) у кластері, обчисленим на основі функції відстані екземплярів до центрів кластера. Подібно до чітких методів у нечіткому кластерному аналізі, критерій якості розбиття є функцією, визначеною на основі попарних відстаней між екземплярами вибірки в просторі ознак. Обчислення таких відстаней для вибірок великих розмірів є обчислювально витратним завданням, а також потребує значних витрат пам'яті для завантаження всієї вибірки спостережень у пам'ять. Крім того, завдання ускладнюється необхідністю попарного переобчислення відстаней між екземплярами, а також необхідністю переобчислення приналежності екземплярів до кластерів.

Чіткі методи кластерного аналізу [1-17, 27-29], очевидно, є більш точними (забезпечують конкретний результат), але в той же час є грубими та менш адаптивними. Нечіткі методи [18-26] дають нечітку оцінку належності екземпляра до кластера і є менш точними (специфічні в оцінці членства), але в той же час вони є більш

адаптивними порівняно з чіткими методами, але також більш витратними за обсягом обчислень і необхідної пам'яті.

Залежно від способу формування розбиття на кластери методи кластерного аналізу можна поділити на неієрархічні [1-17], в яких кластери не підпорядковані, та ієрархічні [27-29], в яких розбиття здійснюється послідовно за формування вкладених кластерів. Фактично неієрархічні методи реалізують пошук у ширину, тоді як ієрархічні методи реалізують пошук у глибину.

Слід також зазначити, що більшість відомих методів кластерного аналізу залежать від набору ознак, заданих користувачем, і не дозволяють оцінити їх значущість. Це призводить до формування надмірного розбиття, збільшення кількості обчислень, а також зменшує можливість сприйняття отриманого розбиття людиною.

Крім того, якщо набір ознак містить взаємопов'язані, повторювані, подібні ознаки або ознаки, які є дискретними щодо розподіленого в часі значення, традиційні методи кластерного аналізу створять надзвичайно складний, надлишковий і неінтерпретовний розподіл, але вони не зможуть ідентифікувати такі ознаки та усунути або використовувати їх більш ефективно.

Тому існує потреба усунути недоліки чітких і нечітких методів шляхом розробки методу нечіткої кластеризації з урахуванням чіткого розбиття та евристики прискорення пошуку.

1.1 Формальна постановка задачі

Уведемо такі позначення: x – вибірка спостережень, x^s – s-й екземпляр вибірки спостережень, x^s_j – значення j-ї ознаки для s-го екземпляра вибірки, s – номер екземпляра, S – кількість екземплярів у вибірці, j – номер ознаки, N – кількість ознак, що характеризують екземпляри вибірки.

Нехай задана вибірка спостережень $x=\{x^s\}$, $s = 1, 2,..., S$, $x^s = \{x^s_j\}$, $j = 1, 2, ..., N$, тоді завдання кластерного аналізу вибірки x полягає у тому, щоб визначити розбиття вибірки x на K підмножин (кластерів, псевдокласів) з параметрами $C=\{C^k\}$:

$$x = \bigcup_{k=1}^{K} \{x^s \mid x^s \in C^k, s = 1,2,...,S\},$$

$$F(x, C) \to opt.$$

Тут C – набір налаштовуваних параметрів, C^k – параметри k-го кластеру, F – критерій якості моделі (задається користувачем), opt – формальне позначення оптимуму.

Як правило, критерій якості має забезпечувати мінімізацію відстаней між екземплярами в межах одного кластера та максимізацію міжкластерних відстаней екземплярів [1, 12-14]. Тут відстань між екземплярами у просторі ознак розглядається як міра їх подібності.

Отже для заданої вибірки x потрібно конструктивно визначити F і знайти для неї оптимальні (або прийнятні) значення C.

1.2 Метод нечіткої кластеризації з гешуванням відстаней

На відміну від більшості методів кластерного аналізу [1-29], які передбачають обчислення відстаней між усіма екземплярами в просторі ознак, пропонується обчислювати геш-відстань від екземплярів до умовного спільного центру координат для кожного екземпляра, замінюючи N-вимірний вектора координат екземпляра на одну координату, а потім визначати відстань між екземплярами в одновимірному просторі. Це дозволить для великих вибірок завантажувати в пам'ять тільки окремі екземпляри (мінімально - по одному по черзі), зменшуючи обсяг обчислень і мінімальний обсяг необхідної пам'яті.

Також, на відміну від традиційних методів кластерного аналізу [1-29], запропоновано розглядати отриману геш-ознаку [30-49] як псевдовихідну ознаку, розбиваючи її на інтервали, які можна порівняти з мітками псевдовиходу класи – кластери. Це дозволить замінити перерахування пар порівнюваних відстаней ознак впорядкованим набором одновимірних координат екземплярів уздовж хеш-осі, таким чином зменшивши кількість обчислень.

Далі, отримавши грубе чітке розбиття екземплярів вибірки, їм пропонується задати розбиття вхідних ознак на нечіткі терми, визначити на їх основі та розбиття екземплярів правила віднесення екземплярів до кластерів.

На відміну від традиційного метричного підходу до кластерного аналізу [1-17], який передбачає використання всього набору початкових ознак, пропонується оцінювати інформативність ознак [50, 51] і нечітких термів [52, 53] і виключити неінформативні терми,

а також неінформативні ознаки, зберігаючи при цьому прийнятний рівень критерію якості.

Потім ми можемо визначити систему нечіткого логічного виведення типу класифікатора Мамдані-Заде, яка у формі нейронечіткої мережі може бути додатково навчена за допомогою методів оптимізації [54-56] для налаштування параметрів функцій належності до нечітких термів і ваг правил, які забезпечують прийнятні значення функціоналу якості кластеризації.

Вищезазначена редукція ознак та термів зменшить складність обчислень, зменшить обсяг пам'яті, складність нейронечіткої системи, зменшить кількість конфігурованих параметрів мережі та, як наслідок, підвищить рівень її узагальнення даних, а також інтерпретовність.

Формально метод побудови нейро-нечіткої мереж для кластерного аналізу даних, який реалізує описані вище ідеї, можна подати наступним чином.

Етап ініціалізації. Задати вибірку спостережень $x=\{x^s\}$ та користувальницькі значення $\varepsilon \geq 0$, $0 < \alpha \leq 1$ та $0 < \beta \leq 1$.

Етап обчислення гешу. Використовуючи один із методів обчислення гешу [30-49], визначити геші $\{x_*^s\}$ для екземплярів вибірки, де x_*^s – геш-значення для екземпляра x^s. Упорядкувати екземпляри вибірки за значеннями по осі-гешу x_*.

Стадія чіткого розбиття простору ознак. Розбити діапазон геш-значень на інтервали, пронумерувавши їх послідовно. Для кожного екземпляра вибірки x^s зафіксувати номер інтервалу $q^s{}_*$, в який він потрапив за геш-значенням x_*^s, як мітку вихідної ознаки – псевдокласу (кластеру) $y^s = q^s{}_*$, де $q^s{}_*$ – номер інтервалу, до якого s-й екземпляр потрапив за геш-значенням; y^s – мітка вихідної ознаки (псевдокласу або кластера) для s-го екземпляра.

Чітке розбиття діапазонів значень ознак на інтервали (вибір термів) можна здійснити різними способами.

Найпростіший спосіб – розділити діапазон значень кожної ознаки на рівну кількість інтервалів, які мають однакову довжину для відповідної ознаки. За такого розбиття простір вхідних ознак буде поділено рівномірною сіткою (гратами), за параметрами інтервалів

якої легко визначити кількість інтервалів, на які розбитий діапазон значень *j*-ї ознаки N_j ($N_j \geq 2$)..

Також доцільно передбачити обмеження $N_j \geq K$. Також бажано, щоб $N_j << S$.. Для заданої кількості інтервалів N_j визначити $L_{j,q}$ – довжину *q*-го інтервалу значень *j*-ї ознаки (j = 1, 2, ..., N; q = 1, 2,..., N_j):

$$L_{j,q} = \frac{x_j^{\max} - x_j^{\min}}{N_j},$$

$$x_j^{\min} = \min_{s=1,2,...,S} \{x_j^s\},$$

$$x_j^{\max} = \max_{s=1,2,...,S} \{x_j^s\},$$

на основі якого можна розрахувати ліву $l_{j,q}$ і праву $r_{j,q}$ межі *q*-го інтервалу *j*-ї ознаки (j = 1, 2, ..., N; q = 1, 2,..., N_j):

$$l_{j,q} = x_j^{\min} + (q-1)L_{j,q},$$

$$r_{j,q} = x_j^{\min} + qL_{j,q}.$$

Для *s*-го екземпляра номер інтервалу *q*, в який він потрапляє за *j*-ю ознакою, визначимо як:

$$q_j^s = 1 + \left\lfloor \frac{x_j^s - x_j^{\min}}{L_{j,1}} \right\rfloor$$

або як

$$q_j^s = \{q \mid l_{j,q} \leq x_j^s \leq r_{j,q}, q = 1,2,...,N_j\}.$$

Оскільки розмірність вибірки $n = NS$, то для забезпечення узагальнюючих властивостей моделі важливо, щоб $3N_jN \leq n$,, тобто $3N_j \leq S$. У результаті можна рекомендувати установку: $K \leq N_j \leq S/3$. Якщо $S/3$ менше K, тоді встановіть $N_j = K$.

У комірки такої сітки в загальному випадку будуть потрапляти екземпляри різних псевдокласів, оскільки не враховується ознака-геш. Також невідома кількість інтервалів, на які необхідно розділити діапазони значень ознак, щоб досягти прийнятної точності апроксимації меж кластерів. Це розбиття буде обчислювально найшвидшим і найпростішим, але воно міститиме невизначеність у виборі кількості інтервалів, а також не дозволить точно виділяти кластери.

Більш складним є розбиття діапазону значень ознак на інтервали, коли на осі кожної ознаки виділяється різна кількість інтервалів і враховується номер псевдокласу, визначений геш-ознакою [30-49]. Для цього потрібно спроектувати мітки екземплярів (номери псевдокласів) одна за одною на вісь *j*-ї ознаки в порядку зростання її значень.

У цьому випадку може виникнути ситуація, коли для одного й того ж значення координати вздовж осі *j*-ї ознаки знайдеться кілька екземплярів з різними мітками псевдокласів. У цьому випадку екземпляри з міткою, яка дорівнює мітці екземпляра з меншою координатою попередньої групи, повинні бути розміщені першими, а екземпляри з міткою, що дорівнює мітці екземпляра з більшою координатою наступної групи, повинні бути розміщені останніми. Після цього необхідно підібрати інтервали значень *j*-ї ознаки $\{<l_{j,q}, r_{j,q}>\}$ так, щоб в межах одного інтервалу були екземпляри з однаковим значенням номера геш-псевдокласу та екземпляри суміжних інтервали значень ознак мали різні геш-номери псевдокласів. Тут можливі ситуації, коли сусідні інтервали можуть стикатися і частково перекриватися, якщо ліва і права межі суміжних інтервалів мають однакові координати. Також можливо, що між сусідніми інтервалами будуть порожнечі, в яких немає екземплярів.

Після формування такого розбиття кількість інтервалів, на які розбивається діапазон значень ознаки, N_j може бути використана для визначення інформативності ознак.

Чим на більше інтервалів зміни номера класу розбивається діапазон ознаки, тим складнішою і нелінійнішою є класифікація, тобто меншою є індивідуальна інформативність цієї ознаки. Чим менше інтервалів значень ознак (в ідеалі, один) відповідає конкретному номеру псевдокласу, тим цінніший цей інтервал для

цього псевдокласу. Кількісно це можна визначити показником індивідуальної інформативності *j*-ї ознаки:

$$I_j = \frac{K}{N_j}, N_j \geq K.$$

Цей показник буде наближатися до нуля при меншій індивідуальній інформативності ознаки і до одиниці – при більшій.

Етап формування правила. Перетворіть кожен екземпляр вибірки у чітке правило в формі:

$$(s): \text{якщо } \bigcup_{j=1}^{N}\left\{\bigcap_{q=1}^{N_j} x_j^s \in [l_{j,q}, r_{j,q}]\right\}, \text{ то } y^s = q^s{}_* \text{ з вагою } w^s{=}1.$$

Тут (*s*) – номер правила, w^s – вага *s*-го правила;

Для спрощення програмної обробки такі правила можна представити у вигляді набору *R*: $\{(s): \{q^s_j\} \rightarrow q^s{}_*, w^s\}$, де q^s_j номер інтервалу, у який *s*-й екземпляр потрапляє за *j*-ю ознакою.

Етап оцінювання якості розбиття та набору правил. Для оцінювання якості згенерованого розбиття та набору правил можна використовувати помилку класифікації *E*.

Для цього для кожного *s*-го екземпляра вибірки x^s, s=1, 2, ..., S:

– визначити його належність до кожного терма $\{q^s_j\}$;

– із множини правил *R* вибрати ті правила, які відповідають розпізнаному екземпляру з лівої сторони (сформувати конфліктну множину правил R_c та оцінити її потужність (кількість правил у конфліктній множині $|R_c|$));

– визначити як обчислений номер класу для розпізнаного екземпляра x^s і конфліктного набору R_c значення:

$$y^s_* = \arg\max_{q=1,2,..,K}\left\{\sum_{p=1}^{|R_c|} w^p\right\}.$$

Для вибірки розпізнаних екземплярів помилку класифікації можна оцінити як:

$$E = \sum_{s=1}^{S} \{1 \mid y^s \neq y_*^s\}.$$

Якщо значення помилки є неприйнятним ($E > \varepsilon$), тоді можна переглянути згенероване розбиття, збільшивши кількість інтервалів, на які розділені діапазони значень ознак та/або змінити метод обчислення гешу.

Етап редукції правил. Усі правила мають бути відсортовані за значеннями $y=\{q^s{}_*\}$, а потім за значеннями $\{q^s{}_j\}$. Встановіть $S' = S$, де S' – кількість екземплярів у скороченій вибірці.

Послідовний переглядаючи правила s= 1, 2, ..., S'–1:: для двох правил (s) та (s+1), послідовних за y, якщо їхні праві частини однакові ($q^s{}_* = q^p{}_*$) та ліві частини однакові ($\forall j = 1,2,....,N : q^s{}_j = q^{s+1}{}_j$), тоді залишити перше ($s$-те) правило, збільшуючи його вагу на вагу (s+1)-го правила: $w^s = w^s + w^{s+1}$, потім видалити друге ((s+1)-ше) правило та зменшити S': $S' = S' - 1$.

Етап скорочення термів і ознак. Для кожного терма кожної вхідної ознаки, використовуючи набір згенерованих правил для кожного k-го псевдокласу, визначити, скільки разів терм використовувався в правилах, враховуючи їхні ваги:

– кількість екземплярів k-го класу, що належать q-му терму j-ї ознаки:

$$N_{j,q,k} = \sum_{s=1}^{S'} \{w^s \mid q_j^s = q, q_*^s = k\};$$

– кількість разів, коли q-й терм j-ї ознаки використовувався в правилах;

$$N_{j,q} = \sum_{s=1}^{S'} \{w^s \mid q_j^s = q\}.$$

Чим більше значення $N_{j,q,k}$, тим сильніше q-й терм j-ї ознаки бере участь у прийнятті рішень про призначення екземпляра k-му псевдокласу.

Визначимо показник індивідуальної інформативності q-го терма j-ї ознаки для всієї множини класів ($j = 1, 2, ..., N$; $q = 1, 2,..., N_j$):

$$I_{j,q} = \frac{\max\limits_{k=1,2,...,K} \{N_{k,j,q}\}}{N_j}.$$

Цей показник буде приймати значення в діапазоні від нуля до одиниці. Чим менше буде його значення, тим менш інформативним буде q-й терм j-ї ознаки. Чим більше буде його значення, тим суттєвішим буде q-й терм j-ї ознаки.

У порядку зростання оцінки індивідуальних інформативностей ознак I_j послідовно для поточної розглянутої j-ї ознаки:

– виключити її (всі її умови) з усіх правил;

– оцінити похибку визначення номера класу для екземплярів вибірки згідно з поточним набором правил і термів E;

– якщо помилка E є прийнятною ($E \le \varepsilon$), тоді вважати j-ту ознаку неінформативною та вилучити її з подальшого розгляду, а також вилучити її терми та функції належності;

– якщо помилка E є неприпустимою ($E > \varepsilon$), тоді повернути видалену ознаку та терми до правил і припинити подальший перегляд ознак.

Переглядаючи терми ознак у порядку від найменш часто використовуваних у правилах до найбільш часто використовуваних (тобто в порядку зростання значення $I_{j,q}$), послідовно для кожного терма:

– виключити його з усіх правил;

– оцінити похибку визначення номера класу для екземплярів вибірки згідно з поточним набором правил і термів E;

– якщо помилка є допустимою ($E \le \varepsilon$), тоді вважати q-й терм j-ї ознаки неінформативним і вилучити його та його функцію належності з подальшого розгляду;

– якщо помилка є неприйнятною ($E > \varepsilon$), тоді повернути видалений термін до правил і припинити подальший перегляд термів.

Етап виявлення подібності та видалення схожих ознак. Для i=1, 2, ..., N, $j = i+1, i+2,...,N$ визначити оцінку попарного зв'язку i-ї та j-ї вхідних ознак $I(i,j) = I(j,i)$. Це можливо реалізувати на основі показників індивідуальної інформативності ознак [48-51], беручи як вхідну ознаку i-ту ознаку, а як вихідну – j-ту ознаку.

На основі попарних оцінок зв'язку між i-ю та j-ю вхідними ознаками $\{I(i,j)\}$ визначити середню оцінку зв'язку ознак:

$$\bar{I} = \frac{1}{0{,}5N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(i,j).$$

Розділити ознаки на групи так, щоб ознаки однієї групи мали оцінку зв'язку, вищу за середню, помножену на визначений користувачем коефіцієнт α, $0 < \alpha \le 1$. Для цього спочатку внести до набору неврахованих ознак усі ознаки, які присутні в правилах поточного набору правил. Потім, поки набір неврахованих ознак непорожній, повторювати:

– вибрати із множини окремо нерозглянутих найбільш інформативну ознаку (по відношенню до вихідної ознаки) і сформувати для неї нову групу ознак;

– із множини неврахованих ознак вибрати всі ознаки, пов'язані з ознакою нової групи сильніше середнього зв'язку з урахуванням коефіцієнта α: $I(i,j) \ge \alpha\bar{I}$, перенести їх до групи цієї ознаки.

З ознак кожної групи, що містить дві або більше ознак:

– залишити в правилах лише одну ознаку, яка найбільше пов'язана з вихідною ознакою (має найвищу оцінку індивідуальної інформативності I_j).

– оцінити похибку класифікації E для поточного набору правил та термів;

– якщо помилка є прийнятною ($E \le \varepsilon$), тоді вилучити виключені ознаки та їх терми з подальшого розгляду, інакше повернути всі ознаки цієї групи.

Альтернативним варіантом є послідовне видалення з кожної групи окремо найменш значущої ознаки (з нижчим значенням I_j), поки помилка залишається прийнятною, а кількість ознак, що залишилися в групі, буде принаймні дорівнювати одиниці.

Крім того, у низці задач, де передбачається, що ряд вхідних ознак є непрямими спостереженнями прихованого фактора або деякі з

вхідних ознак є дискретними відліками розподіленого сигналу, тоді замість або на додаток до цього етапу можна включити не видалення, а поєднання основних ознак у розрахункову. У цьому випадку після виділення груп ознак для кожної групи за всіма її ознаками необхідно розрахувати значення ознак штучної згортки (у найпростішому випадку це може бути сума, середнє значення, максимум, мінімум і добуток значень первинних ознак групи), а потім оцінити для кожної згортки її зв'язок з вихідною ознакою. Якщо жодна згортка не перевищує індивідуальну інформативність первинних ознак групи за значенням індивідуальної інформативності, то для цієї групи слід обмежитися вибором однієї первинної ознаки з найбільшою індивідуальною інформативністю, інакше всі первинні ознаки повинні бути виключити з групи, замінивши їх на штучну ознаку - згортку, визначити терми та їхні параметри для даної ознаки, налаштувавши згенероване розбиття та його параметри, а також набір правил.

Етап виявлення подібності та скорочення подібних термів різних ознак. Визначити оцінку попарної еквівалентності для термів різних вхідних ознак:

$$I_{i,p,j,q} = I_{j,q,i,p} = \frac{\sum_{s=1}^{S} \{1 \mid l_{i,p} \le x_i^s \le r_{i,p}, l_{j,q} \le x_j^s \le r_{j,q}\}}{\max\{N_{i,p}, N_{j,q}\}},$$
$$i = 1, 2, ..., N, j = i+1, i+2, ..., N,$$
$$p = 1, 2, ..., N_i, q = 1, 2, ..., N_j.$$

Визначити середню оцінку зв’язку між складовими різних вхідних ознак:

$$\bar{I}^t = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{p=1}^{N_i} \sum_{q=1}^{N_j} I_{i,p,j,q}}{0{,}5N(N-1)\left(\sum_{i=1}^{N} N_i\right)^2}.$$

Розділіть терми на групи так, щоб терми різних ознак однієї групи мали оцінку зв'язку більшу, ніж середнє значення, помножене на заданий користувачем коефіцієнт β, $0 < \beta \le 1$.

Для цього спочатку в набір неврахованих термінів внести усі терми, які присутні в правилах поточного набору правил. Потім, поки множина нерозглянутих термінів непорожня, повторювати:

– вибрати із множини окремо нерозглянутих найбільш інформативний терм і сформувати для нього нову групу термів;

– із множини нерозглянутих термів, виключаючи інші терми ознаки, такої ж як ознака групоутворюючого терма, вибрати всі терми, які пов'язані з термом нової групи сильніше середнього зв'язку, враховуючи коефіцієнт β: $I_{i,p,j,q} \ge \beta \bar{I}^{t}$, перенести їх до групи цього терма.

З термв кожної групи, що містить два або більше термів:

– залишити в правилах лише один терм, ознака якого найбільше пов'язана з вихідною ознакою (має найбільше значення індивідуальної інформативності I_j);

– оцінити помилку класифікації E для поточного набору правил і термів;

– якщо помилка є прийнятною ($E \le \varepsilon$), тоді вилучити виключені терми з подальшого розгляду, інакше – повернути всі терми цієї групи.

Альтернативним варіантом є послідовне видалення з кожної групи окремо найменш значущого терма (з меншим значенням I_j), доки помилка залишається прийнятною, а кількість термів, що залишилися в групі, буде принаймні дорівнювати одному.

Етап формування нечітких термів. На основі параметрів інтервалів значень ознак, відібраних при формуванні чіткого розбиття, і термів та ознак, відібраних у процесі редукції, можна визначити функції належності до нечітких термів. Для цього можна використовувати різні типи елементарних функцій належності [21, 22]. Для чітких інтервалів, в які потрапили два або більше екземплярів, пропонується використовувати такі функції: трапецієподібну, дзвоноподібну, гауссову, П -подібну. Для точкових інтервалів, куди потрапив лише один екземпляр, пропонується використовувати такі функції: трикутну, дзвоноподібну, гаусову, П-подібну. Кожна з цих функцій

$\mu_{j,q}(x_j^s)$ для певного нечіткого терма (q-го інтервалу значень на осі j-ї ознаки) матиме параметри $<a_{j,q}, b_{j,q}, c_{j,q}, d_{j,q}>$, такі, що: $a_{j,q} \le b_{j,q} \le c_{j,q} \le d_{j,q}$. Значення параметрів функцій належності можуть бути визначені на основі параметрів чіткого розбиття.

Наприклад, для трапецієподібної та П-подібної функцій параметри можна визначити як:

– для розбиття на однакові за довжиною інтервали:

$$a_{j,q} = b_{j,q} = l_{j,q},$$

$$c_{j,q} = d_{j,q} = r_{j,q};$$

– для розбиття на інтервали різних класів:

$$a_{j,q} = \begin{cases} l_{j,q}, q = 1; \\ \dfrac{r_{j,q-1} + l_{j,q}}{2}, q > 1; \end{cases}$$

$$b_{j,q} = l_{j,q},$$

$$c_{j,q} = r_{j,q},$$

$$d_{j,q} = d_{j,q} = \begin{cases} r_{j,q}, q = N_j; \\ \dfrac{r_{j,q} + l_{j,q+1}}{2}, q < N_j. \end{cases}$$

Етап формування нейро-нечіткої мережі для кластеризації. Відобразимо згенеровану базу знань у систему нечіткого логічного виведення, яку зручно представимо як нейронну мережі.

Структуру мережі можна визначити на основі апроксиматора Мамдані-Заде [57]. Вузли вхідного шару мережі відповідатимуть вхідним ознакам x_j для розпізнаного екземпляра $x^s = \{x_j^s\}$. Таким

чином, вхідний шар матиме N вузлів (тут і далі маються на увазі не початкові значення кількості ознак і термів, а після скорочення).

Вузли першого прихованого шару мережі будуть відповідати блокам фаззіфікації, тобто визначати значення функцій належності для термів вхідних ознак. На першому прихованому шарі буде $\sum_{j=1}^{N} N_j$ вузлів. Вхід кожного вузла першого прихованого шару отримує значення з виходу лише вузла вхідного шару, що відповідає його ознаці.

Вузли наступного другого прихованого шару об'єднують функції належності термів у функції приналежності антецедентів (лівих частин) правил, об'єднавши виходи вузлів першого шару, відповідні терми яких включені в відповідні антецеденти. Другий прихований шар матиме S' нечітких вузлів «ТА».

Вузли третього шару будуть об'єднувати правила в псевдокласи, реалізуючи нечітке «АБО». Третій прихований шар матиме K вузлів.

Єдиний вузол вихідного шару дефаззіфікує результат, видаючи номер кластеру (псевдокласу) за формулою:

$$y^s = \left\lfloor \frac{\sum_{k=1}^{K} k\mu_k(x^s)}{\sum_{k=1}^{K} \mu_k(x^s)} \right\rfloor$$

або

$$y^s = \arg \max_{k=1,2,\ldots,K} \{\mu_k(x^s)\}.$$

Параметри мережі будуть визначені на основі попередньо сформованого чіткого розбиття та набору правил.

Етап додаткового навчання та оптимізації нейронечіткого кластеризатора. Оцінимо якість продуктивності нейро-нечіткої мережі (якість кластеризації даних) на основі заданого функціоналу F, який можна визначити на основі широкого класу метрик [1, 2, 7, 8, 12-15]. За допомогою методів еволюційної оптимізації [54, 55] можна підібрати такі значення параметрів функцій належності до термів

мережі, які покращать значення оптимізованого функціонала *F*. Кінцевою моделлю буде нейронечіткий кластеризатор, оптимізований за кількістю використовуваних ознак і функціоналом *F*.

1.3 Експериментальне дослідження методу кластеризації

Для дослідження практичної застосовності запропонованого методу його було реалізовано програмно та використано для вирішення комплексу практичних задач різного характеру та розмірності. Характеристики вибірок вихідних даних практичних завдань наведено в табл. 1.1.

Таблиця 1.1 – Характеристики вихідних вибірок для кластерного аналізу

Назва задачі	Абревіатура задачі	Джерело	*N*	*S*	*n*
Low Resolution Spectrometer	LRS	https://archive.ics.uci.edu/ml/datasets/Low+Resolution+Spectrometer	102	531	54162
Musk (Version 2)	Mv2	https://archive.ics.uci.edu/ml/datasets/Musk+%28Version+2%29	168	6598	1108464
Urban Land Cover	ULC	https://archive.ics.uci.edu/ml/datasets/Urban+Land+Cover	148	168	24864
Iris	IRIS	https://archive.ics.uci.edu/ml/datasets/Iris	4	150	600
Heart Disease	HD	https://archive.ics.uci.edu/ml/datasets/Heart+Disease	75	303	22725
Breast Cancer Wisconsin (Diagnostic)	BCWD	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29	32	569	18208
Arrhythmia	ART	https://archive.ics.uci.edu/ml/datasets/Arrhythmia	279	452	126108
Crop mapping using fused optical-radar	CMFOR	https://archive.ics.uci.edu/ml/datasets/Crop+mapping+using+fused+optical-radar+data+set	175	325834	57020950
Sensorless Drive Diagnosis	SDD	https://archive.ics.uci.edu/ml/datasets/Dataset+for+Sensorless+Drive+Diagnosis	49	58509	2866941

Для кожноъ задачі в експериментах використовувалися різні методи генерації хешів і різні значення параметрів, що регулюють роботу запропонованого методу.

Для оцінювання результатів методів кластерного аналізу використовувалися: похибка *E*, значення *F* для результуючої моделі, час роботи методу *t*, обсяг пам’яті, використаний методом, *m*,

кількість параметрів кластеризувальної моделі кластеризації N_w та кількість ознак після редукції N'. Крім описаних вище характеристик, пропонується також використовувати такі показники.

Узагальнюючі властивості отриманих моделей кластеризації порівняно з вихідною розмірністю даних можна охарактеризувати коефіцієнтом узагальнення:

$$I_G = \frac{NS}{N_w} = \frac{n}{N_w}, N_w \geq 1.$$

Цей коефіцієнт прийматиме значення в діапазоні від нуля до n. Чим більше параметрів матиме модель, тим нижче буде рівень узагальнення відносно розмірності вихідних даних, тим менше буде значення коефіцієнта узагальнення.

Альтернативно, узагальнення може бути охарактеризовано в однаковій кількості випадків як відношення кількості ознак у первинному наборі N до кількості ознак, використаних у скороченому наборі кінцевої моделі, N':

$$I_{GF} = \frac{N}{N'}, N \geq N' \geq 1.$$

Цей коефіцієнт прийматиме значення в діапазоні від нуля до N. Чим більше параметрів матиме модель, тим нижче буде рівень узагальнення відносно розмірності вихідних даних, тим менше буде значення коефіцієнта узагальнення.

При попарному порівнянні отриманих моделей 1 і 2 для однієї і тієї ж вихідної вибірки даних їх узагальнення з допустимими значеннями критерію F можна охарактеризувати співвідношеннями:

$$G_{1,2} = \frac{I_{G1}}{I_{G2}} = \frac{NS}{N_{w1}} \frac{N_{w2}}{NS} = \frac{N_{w2}}{N_{w1}}, N_{w1} \geq 1, N_{w2} \geq 1, N \geq 1, S \geq 1,$$

$$GF_{1,2} = \frac{I_{GF1}}{I_{GF2}} = \frac{N}{N'_1} \frac{N'_2}{N} = \frac{N'_2}{N'_1}, N \geq 1, N'_1 \geq 1, N'_2 \geq 1,$$

де $F_{1,2}$ – частка від ділення F для першої та другої моделей, $G_{1,2}$ – частка від ділення коефіцієнтів I_G для першої та другої моделей, $GF_{1,2}$ – частка від ділення коефіцієнтів GF для першої та другої моделей.

Позначимо: $t_{1,2}$ – частка від ділення t для першого та другого методів, $m_{1,2}$ – частка від ділення m для першого та другого методів.

Для двох моделей 1 і 2 ми також можемо визначити:

$$t_{1,2} = t_1 / t_2,$$

$$m_{1,2} = m_1 / m_2,$$

$$F_{1,2} = F_1 / F_2.$$

У таблиці 1.2 наведено значення показників $G_{1,2}$, $GF_{1,2}$, $t_{1,2}$, $m_{1,2}$, та $F_{1,2}$ для порівняння запропонованого методу з методом нечітких с-середніх [24-26], в якому початкове розбиття формується випадковим чином.

Таблиця 1.2 – Результуючі значення показників для порівняння кластеризувальних моделей

Абревіатура задачі	$GF_{1,2}$	$G_{1,2}$	$t_{1,2}$	$m_{1,2}$	$F_{1,2}$
LRS	1,3	1,5	0,8	0,7	0,9
Mv2	1,4	2,4	0,9	0,4	1,1
ULC	1,5	1,7	0,9	0,6	0,9
IRIS	1,3	1,4	1	0,8	1
HD	1,2	1,5	0,9	0,6	1,1
BCWD	1,1	1,7	1,1	0,6	0,9
ART	1,3	2,2	0,9	0,5	0,9
CMFOR	1,2	1,9	0,8	0,5	1,1
SDD	1,2	1,7	0,9	0,6	0,9

З табл. 1.1 і Табл. 1.2 легко побачити, що запропонований метод дозволяє для однієї і тієї ж вибірки даних значно покращити узагальнюючі властивості моделі, скоротити витрати часу та пам’яті комп’ютера, а також забезпечити краще або прийнятне значення функціоналу якосты. Це пояснюється тим, що запропонований метод невипадково генерує розбиття простору ознак, відбирає та скорочує

неінформативні терміни та ознаки, прагнучи зменшити складність моделі. При цьому запропонований метод не вимагає обчислення відстаней між екземплярами у *N*-вимірному просторі ознак завдяки використанню локально чутливого хешу.

Отримані експериментально узагальнені залежності між основними характеристиками запропонованого методу схематично наведено на рис. 1.1-рис. 1.9.

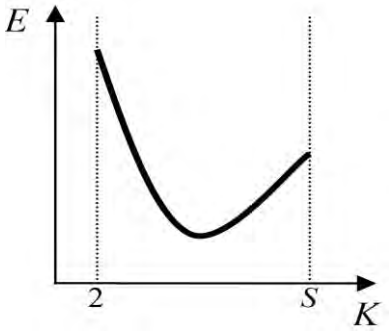


Рисунок 1.1 – Схематичний графік усередненої залежності *E* від *K*

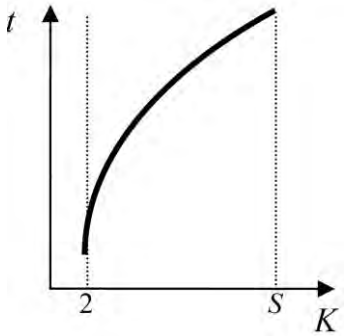


Рисунок 1.2 – Схематичний графік усередненої залежності *t* від *K*

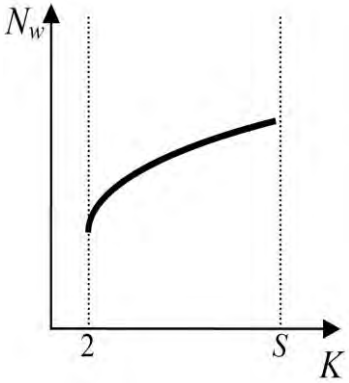


Рисунок 1.3 – Схематичний графік усередненої залежності N_w від K

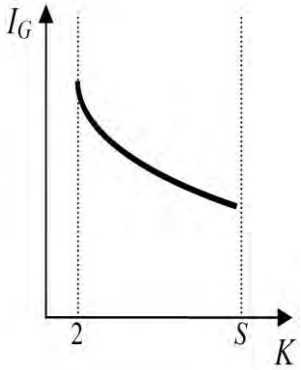


Рисунок 1.4 – Схематичний графік усередненої залежності I_G від K

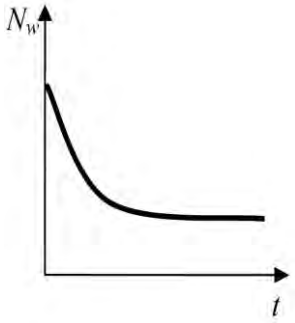


Рисунок 1.5 – Схематичний графік усередненої залежності N_w від t

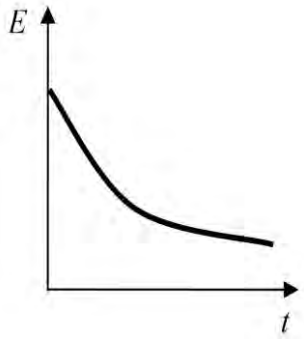


Рисунок 1.6 – Схематичний графік усередненої залежності E від t

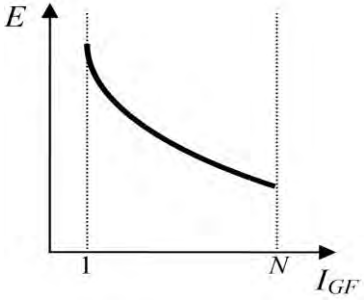


Рисунок 1.7 – Схематичний графік усередненої залежності E від I_{GF}

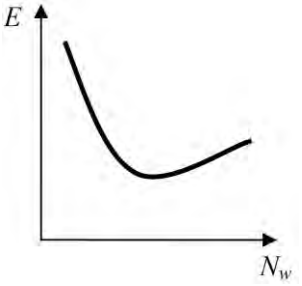


Рисунок 1.8 – Схематичний графік усередненої залежності E від N_w

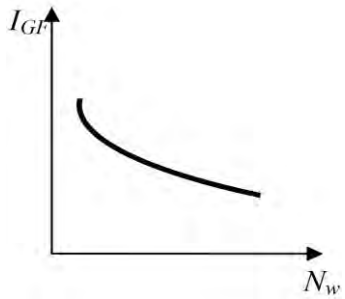


Рисунок 1.9 – Схематичний графік усередненої залежності I_{GF} від N_w

Як видно з рис. 1.1, зі збільшенням значення K значення похибки E зменшується до певного рівня, після чого починає зростати. Зменшення похибки E зі збільшенням K пояснюється підвищенням точності апроксимації меж кластерів за рахунок уточнення розбиття вихідної ознаки. Збільшення похибки Е при подальшому збільшенні значення K пояснюється збільшенням невизначеності апроксимації меж кластерів із надмірно детальним розбиттям вихідної ознаки внаслідок виділення великої кількості дрібних кластерів. Тому рекомендується в процесі виконання методу ітераційно вибирати таке значення K при використанні рівномірного розбиття простору ознак, при якому буде досягнуто найменше значення помилки. Використання нерівномірного розбиття на інтервали з різними номерами класів дозволяє автоматизувати цей процес і уникнути зростання похибки. Як правило, зі збільшенням кількості кластерів точність розбиття зростає, але вартість обчислень і вимоги до пам’яті також зростають.

Як видно з рис. 1.2 та рис. 1.3, зі збільшенням кількості псевдокласів K значно збільшуються час роботи методу t та кількість параметрів результуючої моделі N_w. Це пояснюється тим, що зі збільшенням кількості псевдокласів (псевдокластерів) K буде зростати кількість виділених термів та їх параметрів, що вимагатиме значного часу для розрахунку їх показників інформативності та редукції, а також для оптимізаційного налаштування збільшеної кількості параметрів моделі нейро-нечіткої мережі.

Як видно з рис. 1.4, зі збільшенням кількості псевдокластерів K спостерігається зменшення показника узагальнення моделі I_G. Це пояснюється значним збільшенням кількості параметрів моделі за

рахунок більш детальної апроксимації меж кластерів за рахунок збільшення їх кількості.

Як видно з рис. 1.5, зі збільшенням часу, витраченого на процес роботи методу t, кількість параметрів моделі N_w зменшується. Це пояснюється тим, що кількості ознак і термів скорочуються за рахунок видалення неінформативних і дублюючих ознак і термів. Чим більше ітерацій буде в процесі виявлення та редукції неінформативних термів і ознак, тим більше часу витратить метод, але тим меншою буде кількість параметрів результуючої моделі.

Як видно з рис. 1.6, зі збільшенням часу роботи методу t спостерігається зменшення похибки моделі E. Це пояснюється тим, що коригування параметрів моделі дає можливість підвищити точність (зменшити похибку) моделі. Також збільшення витрат часу можна пояснити ітеративним пошуком оптимального розбиття простору ознак, що в кінцевому підсумку призведе до зменшення похибки результуючої моделі.

Як видно з рис. 1.7, зі збільшенням значення показника I_{GF} спостерігається зменшення похибки моделі E. Це пояснюється тим, що при дуже високому узагальненні кількість використовуваних ознак буде меншою і, відповідно, більш грубою буде апроксимація розбиття простору ознак, що призведе до збільшення похибки E. Чим нижче буде значення індексу узагальнення ознак, тим буде використано більше ознак, а простір ознак буде поділено більш детально, що зменшить значення помилки.

Як видно з рис. 1.8, зі збільшенням кількості параметрів моделі N_w відбувається падіння значення похибки E до певного значення. Це пояснюється тим, що деталізація розбиття простору ознак за рахунок збільшення кількості псевдокластерів дає змогу більш точно апроксимувати межі кластерів. Подальше зростання похибки E в процесі збільшення значення N_w пояснюється тим, що надмірний відбір кластерів призводить до відсутності узагальнення, що поступово відображається на зростанні значення похибки E. Однак це є типовим в основному для рівномірного розбиття простору ознак.

Як видно з рис. 1.9, зі збільшенням кількості параметрів моделі N_w, спостерігається зменшення індексу узагальнення ознак I_{GF}. Це пояснюється тим, що деталізація поділу ознакового простору призводить до формування більшої кількості неінформативних термів і ознак, що дає змогу виключити неінформативні ознаки. З іншого

боку, збільшення кількості параметрів моделі N_w можна пояснити збільшенням кількості ознак у вихідному наборі ознак N для задачі, що, у свою чергу, може вказувати на більшу частку неінформативних ознак, тобто про зменшення кількості ознак.

На рис. 1.10 наведено схематичні графіки усереднених залежностей E, I_G, t та N_w від значень α та β.

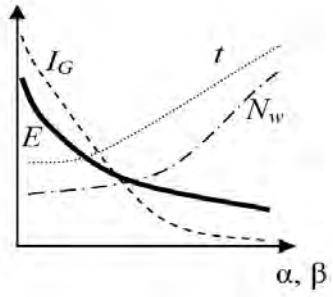


Рисунок 1.10 – Схематичні графіки усереднених залежностей E, I_G, t та N_w від значень α та β

Як видно з рис. 1.10, чим більше буде значення α або β, тим быльше будуть значення t та N_w і тим нижче будуть значення E та I_G. Якщо передбачається, що ознаки мають різну природу, то значення коефіцієнтів α та β рекомендується встановлювати більшими. Якщо передбачається, що ознаки є впорядкованими відліками певної величини, то значення коефіцієнтів α та β рекомендується встановлювати меншими.

В результаті аналізу експериментально отриманих даних підтверджено працездатність і практичну застосовність запропонованого методу та розробленого програмного забезпечення.

Запропонований метод поєднує в собі ідеї чіткого та нечіткого кластерного аналізу. Спочатку він створює чітке розбиття простору ознак, але потім, завдяки фаззіфікації, перетворює його на нечітке розбиття. Чітке розбиття використовується для автоматизації вибору кількості кластерів, а також для прискорення вибору термів ознак. Запропонований метод генерує чітке розбиття автоматично (без втручання людини), він є швидшим, ніж традиційне ітераційне

(оптимізаційне) формування нечіткого розбиття [13, 21-23], яке властиве більшості методів нечіткої кластеризації.

На відміну від традиційно використовуваних метричних методів кластерного аналізу [1-3], які передбачають використання всього первинного набору ознак у кінцевій моделі, запропонований метод вибирає мінімальну підмножину ознак, необхідну для кластеризації, тим самим зменшуючи структурну та параметричну складності моделі, підвищуючи її узагальнювальні властивості та інтерпретабельність (пояснюваність), а також може зменшити кількість ознак шляхом комбінування первинних ознак у штучно розраховані, тим самим додатково збільшуючи узагальнювальні властивості моделі та її інтерпретабельність. Крім того, запропонований метод є більш адаптивним за рахунок фаззіфікації, не вимагає початкового налаштування кількості кластерів і початкового розбиття вибірки на кластери, а також налаштування користувачем метрик для кластерів.

На відміну від ієрархічних методів кластерного аналізу [27-29], які підпорядковують ознаки та утворюють ієрархію розбиття, яка може мати велику глибину, запропонований метод не підпорядковує ознаки, але при цьому видаляє неінформативні ознаки та терми, а ієрархія його глибинних перевірок не перевищує трьох рівнів. При цьому запропонований метод значно перевищує ієрархічні методи за розпаралеленням обчислень, що досягається за рахунок меншої глибини порівняно з ієрархічними методами. Проте запропонований метод дає змогу отримати як додатковий результат оцінки інформативності ознак і термів, формувати штучні ознаки шляхом заміни вихідних, адаптивно коригувати форму та параметри кластерів за рахунок функцій належності, а також автоматично генерувати кількість кластерів.

1.4 Висновки за розділом 1

Розглянуто задачу кластерного аналізу багатовимірних даних. У рамках цієї задачі підвищено швидкість формування кластерів, зменшено складність кластеризувальної моделі та підвищено її інтерпретабельність.

Наукова новизна отриманих результатів полягає в тому, що вперше запропоновано метод кластерного аналізу багатовимірних даних, який

для кожного екземпляра розраховує свій геш на основі відстані до умовного центру координат, використовує одновимірну координату за віссю гешу для визначення відстаней між екземплярами, розглядає результуючий геш як псевдовихід ознаки, розбиваючи його на інтервали, з якими порівнює мітки псевдокласів-кластерів, отримуючи грубе чітке розділення простору ознак і екземплярів вибірки, автоматично генерує розбиття вхідних ознак на нечіткі терми, визначає правила віднесення екземплярів до кластерів і, як результат, формує систему нечіткого виведення класифікатора Мамдані-Заде, яка донавчається у формі нейронечіткої мережі для забезпечення прийнятного значення функціоналу якості кластеризації. Це дає змогу зменшити кількість використовуваних термів і ознак, оцінити їх внесок у прийняття рішень про призначення екземплярів кластерам, збільшити швидкість кластерного аналізу даних і підвищити інтерпретативність розбиття результуючих даних на кластери.

Практичне значення отриманих результатів полягає в тому, що розроблено математичне забезпечення, яке дозволяє вирішити задачу кластерного аналізу даних в умовах великої розмірності даних. Проведені експерименти підтвердили працездатність розробленого програмного забезпечення. Вони дозволяють рекомендувати його для використання на практиці в задачах аналізу даних різної природи та розмірності.

Перспективи подальших досліджень полягають у вивченні застосування запропонованого методу на широкому спектрі практичних задач різного розміру та характеру, у дослідженні впливу різноманітних метрик на результати методу (точність та швидкість побудови нейро-нечітких мереж, обчислювальна складність), розробити паралельну реалізацію методу, вивчити питання інтеграції методу з еволюційними та мультиагентними методами пошуку.

Результати, наведені у розділі 1, опубліковані в [58].

1.5 Література до розділу 1

1. Everitt B. Cluster analysis / [B. Everitt, S. Landau, L. Morven et al.]. – Chichester: Wiley, 2011. – 330 p.

2. Data Clustering : Algorithms and Applications / eds.: C. Aggarwal, C. Reddy, K. Chandan. – New York: Chapman and Hall/CRC, 2016. – 652 p.
3. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values / Z. Huang // Data Mining and Knowledge Discovery. – 1998. – Vol. 2, Issue 3. – P. 283–304. DOI:10.1023/A:1009769707641.S2CID 11323096.
4. Ng R. Efficient and effective clustering method for spatial data mining / R. Ng, J. Han // 20th International Conference on Very Large Data Bases (VLDB'94), September 12-15, 1994, Santiago, Chile : proceedings. –Burlington: Morgan Kaufmann,1994. – P. 144–155.
5. Bailey K. D. Typologies and Taxonomies: An Introduction to Classification Techniques / K. D. Bailey. – London: Sage Publications, 1994. – 96 p.
6. Gordon A.D. Classification / A.D. Gordon. – Boca Raton: Chapman & Hall/CRC, 1999. – 256 p.
7. Romesburg C.H. Cluster Analysis for Researchers / C.H. Romesburg. – Belmont : Lifetime Learning Publications, 1984. – 334 p.
8. Aldenderfer M.S. Cluster Analysis / M.S Aldenderfer, R. K. Blashfield. – London : Sage Publications, 1984. – 88 p.
9. Meilă, M. Comparing Clusterings by the Variation of Information / M. Meilă // Lecture Notes in Computer Science. –2003. – Vol. 2777. – P. 173–187. DOI:10.1007/978-3-540-45167-9_14.
10. Hierarchical Clustering Based on Mutual Information / [A. Kraskov, H. Stögbauer, R.G. Andrzejak, P. Grassberger, Peter] : [Electronic resource]. – Access mode: https://arxiv.org/abs/q-bio/0311039.
11. Frey B. J. Clustering by Passing Messages Between Data Points / B. J. Frey, D. Dueck // Science. – 2007. –V. 315, № 5814. – P. 972–976. DOI: 10.1126/science.1136800.
12. Pfitzner D. Characterization and evaluation of similarity measures for pairs of clusterings / D. Pfitzner, R. Leibbrandt, D. Powers // Knowledge and Information Systems. –2009. – Vol. 19, № 3. – P. 361–394. DOI:10.1007/s10115-008-0150-6.
13. Dunn J. Well separated clusters and optimal fuzzy partitions / J. Dunn // Journal of Cybernetics. – 1974. – № 4. – P. 95–104. DOI:10.1080/01969727408546059.

14. Rand W. M. Objective criteria for the evaluation of clustering methods / W. M. Rand // Journal of the American Statistical Association. – 1971. – Vol. 66 (336). – P. 846–850. DOI:10.2307/2284239.
15. Hubert L. Comparing partitions / L. Hubert, P. Arabie // Journal of Classification. 1985. – Vol. 2. – P. 193–218. DOI:10.1007/BF01908075.S2CID189915041
16. Di Marco A. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction / A. Di Marco, R. Navigli // Computational Linguistics. – 2013. – Vol. 39, № 3. – P. 709–754. DOI:10.1162/COLI_a_00148. S2CID 1775181.
17. Arnott R. D. Cluster Analysis and Stock Price Comovement / R. D. Arnott // Financial Analysts Journal. –1980. –Vol. 36, № 6. – P. 56–62. DOI: 10.2469/faj.v36.n6.56. ISSN 0015-198X.
18. Dunn J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters / J. C. Dunn // Journal of Cybernetics. – 1973. – Vol. 3, issue 3. – P. 32–57. DOI:10.1080/01969727308546046.
19. A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data / [M. Ahmed S. Yamany, N. Mohamed, A. Farag, T. Moriarty] // IEEE Transactions on Medical Imaging. – 2002. – Vol. 21, № 3. – P. 193–199. DOI:10.1109/42.996338.
20. Abonyi J. Cluster Analysis for Data Mining and System Identification / J. Abonyi, B. Feil. – Berlin: Birkhäuser Verlag, 2007. – 306 p. DOI: 10.1007/978-3-7643-7988-9
21. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition / [F. Höppner, F. Klawonn, R. Kruse, T. Runkler]. – Chichester: John Wiley & Sons, 1999. – 304 p.
22. Miyamoto S. Fuzzy Sets in Information and Retrieval and Cluster Analysis / S. Miyamoto. – Dordrecht: Kluwer Academic Publishers, 1990. – 274 p.
23. Banerjee T. Day or Night Activity Recognition From Video Using Fuzzy Clustering Techniques / T. Banerjee // IEEE Transactions on Fuzzy Systems. – 2014. – Vol. 22, issue 3. – P. 483–493. DOI:10.1109/TFUZZ.2013.2260756.
24. Advances in Fuzzy Clustering and its Applications / eds.: J. Valente de Oliveira, W. Pedrycz. – Chichester: John Wiley & Sons, 2007. – 454 p.

25. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – New York: Plenum Press, 1981. – 272 p. DOI: 10.1007/978-1-4757-0450-1
26. Dumitrescu D. Fuzzy Sets and Applications to Clustering and Training / D. Dumitrescu, B. Lazzerini, L.C. Jain. – Boca Raton: CRC Press, 2000. – 664 p.
27. Finding Hierarchies of Subspace Clusters / [E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, I. Müller-Gorman, A. Zimek] // Lecture Notes in Computer Science. – 2006. – Vol. 4213. – P. 446–453. DOI:10.1007/11871637_42. ISBN 978-3-540-45374-1.
28. Detection and Visualization of Subspace Cluster Hierarchies / [E. Achtert, C. Böhm, H. P. Kriegel, P. Kröger, I. Müller-Gorman, A. Zimek] // Lecture Notes in Computer Science. –2007. – Vol. 4443. – P. 152–163. DOI:10.1007/978-3-540-71703-4_15.
29. Johnson S. C. Hierarchical clustering schemes / S. C. Johnson // Psychometrika. – 1967. – Vol. 32, № 3. – P. 241–254. DOI:10.1007/BF02289588
30. A Survey on Locality Sensitive Hashing Algorithms and their Applications [Electronic resource] / [O. Jafari, P. Maurya, P. Nagarkar, K. M. Islam, C. Crushev]. – Access mode: https://arxiv.org/pdf/2102.08942
31. Buhler J. Efficient large-scale sequence comparison by locality-sensitive hashing / J. Buhler // Bioinformatics. – 2001. – Vol. 17, № 5. – P. 419–428.
32. Zhao K. Locality Preserving Hashing / K. Zhao, H. Lu, J. Mei // Twenty-Eighth AAAI Conference on Artificial Intelligence, 27-31 July 2014, Québec : proceedings. – Palo Alto: AAAI Press, 2014. – P. 2874–2880.
33. Tsai Y.-H. Locality preserving hashing / Y.-H. Tsai, M.-H. Yang // 2014 IEEE International Conference on Image Processing (ICIP), Paris, 27–30 of October 2014: proceedings. – Los Alamitos: IEEE, 2014. – P. 2988–2992. DOI: 10.1109/ICIP.2014.7025604.
34. Feature Hashing for Large Scale Multitask Learning / [K. Weinberger, A. Dasgupta, J. Langford, et al.] // 26th Annual International Conference on Machine Learning (ICML '09) Montreal, June 2009 : proceedings. – New York : ACM, 2009. – P. 1113–1120. DOI: 10.1145/1553374.1553516

35. Wolfson H. J. Geometric Hashing: An Overview / H. J. Wolfson, I. Rigoutsos // IEEE Computational Sci-ence and Engineering. – 1997. – Vol. 4, No 4. – P. 10–21.
36. Fast supervised discrete hashing / [J. Gui, T. Liu, Z. Sunet et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 40, No 2. – P. 490–496. DOI: 10.1109/TPAMI.2017.2678475
37. Aluç, G. Building self-clustering RDF databases using Tunable-LSH / G. Aluç, M. Özsu, K. Daudjee // The VLDB Journal. – 2018. – Vol. 28, № 2. – P. 173–195. DOI:10.1007/s00778-018-0530-9
38. Pauleve L. Locality sensitive hashing: A comparison of hash function types and querying mechanisms / L. Pauleve, H. Jegou, L. Amsaleg // Pattern Recognition Letters. – 2010. – Vol. 31, № 11. – P. 1348–1358. DOI:10.1016/j.patrec.2010.04.004.
39. Andoni A. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions / A. Andoni, P. Indyk // Communications of the ACM. – 2008. – Vol. 51, № 1. – P. 117–122. DOI:10.1145/1327452.1327494.
40. Salakhutdinov R. Semantic hashing / R. Salakhutdinov, G. Hinton // International Journal of Approximate Reasoning. –2008. – Vol. 50, № 7. – P. 969–978. DO:10.1016/j.ijar.2008.11.006.
41. Fast similarity sketching /[S. Dahlgaard, M. Knudsen, M. Thorup] // 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), 15-17 October 2017, Berkeley. – Los Alamitos: IEEE, 2017. – 663-671. DOI: 10.1109/FOCS.2017.67
42. Chin A. Locality-preserving hash functions for general purpose parallel computation / A. Chin // Algorithmica. – 1994. – Vol. 12, issue 2–3. – P. 170–181. DOI:10.1007/BF01185209. S2CID 18108051.
43. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2014. – № 1. – С. 120–128. DOI: 10.15588/1607-3274-2014-1-17
44. Subbotin S. A. The neuro-fuzzy diagnostic model synthesis with hashed transformation in the sequence and parallel mode/ S. A. Subbotin, A. Yu. Blagodarev, Ye. A. Gofman // Radio Electronics, Computer Science, Control. – 2017. – № 1. – С. 56-65. DOI: 10.15588/1607-3274-2017-1-7

45. Subbotin S. A. The polar coordinates based hashing for data dimensionality reduction / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2020. – № 4. – P. 118-128. DOI: 10.15588/1607-3274-2020-4-12
46. Subbotin S. A. Analiz preobrazovaniy dlya proyetsi-rovaniya dannykh na obobshchonnuyu os' v zadachakh raspoznavaniya obrazov / S. A. Subbotin, A. A. Oleynik // Shtuchniy íntelekt. – 2010. – № 1. – C. 114–121.
47. Subbotin S. A. Constructed features for automatic classification of stationary timing signals / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2012. – № 1. – C. 96–103. DOI: 10.15588/1607-3274-2012-1-19
48. Subbotin S. A. The complex data dimensionality reduction for diagnostic and recognition model building on precedents/ S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2016. – № 4. – C. 70-76. DOI: 10.15588/1607-3274-2016-4-9
49. Subbotin S. A. Evaluation of informativity and selection of instances based on hashing / S. A. Subbotin / Radio Electronics, Computer Science, Control. – 2020. – № 3. – C. 129–137. DOI: 10.15588/1607-3274-2020-3-12
50. Oliinyk A. The system of criteria for feature informativeness estimation in pattern recognition / A. Oliinyk, S. Subbotin, V. Lovkin, O. Blagodariov, T. Zaiko // Radio Electronics, Computer Science, Control. – 2017. – № 4. – C. 85–96. DOI: 10.15588/1607-3274-2017-4-10
51. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis / S. Subbotin // Applications of Computational Intelligence in Biomedical Technology / eds.: R. Bris, J. Majernik, K. Pancerz, E. Zaitseva. – Cham : Springer, 2016: – P. 215–228. – (Studies in Computational Intelligence, Vol. 606).
52. Subbotin S. The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition / S. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2013. – Vol. 22, № 2. – P. 97–103.
53. Subbotin S. The Fully-Defined Neuro-Fuzzy Model Synthesis / S. Subbotin, A. Oliinyk // Data Stream Mining & Processing (DSMP):

2016 IEEE First International Conference, Lviv, 23-27 August 20016 : proceedings. – Lviv: NU "Lvivska Politeckhnika", 2016. – P. 9-14.
54. Oliinyk A. O. Using parallel random search to train fuzzy neural networks / A. O. Oliinyk, S. Yu. Skrupsky, S. A. Subbotin // Automatic Control and Computer Sciences. – 2014. – Vol. 48. – №. 6. – P. 313–323.
55. Subbotin S. The method of a structural-parametric synthesis of neuro-fuzzy diagnostic model based on the hybrid stochastic search / S. Subbotin // The experience of designing and application of CAD systems in microelectronics : XI International conference CADSM–2011, Polyana–Svalyava, 23–25 February 2011 : proceedings. – Lviv : NU “Lviv Polytechnic”, 2011. – P. 248 – 249.
56. Subbotin S. A. Building a fully defined neuro-fuzzy network with a regular partition of a feature space based on large sample / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2016. – № 3. – P. 47-53. DOI: 10.15588/1607-3274-2016-3-6
57. Halgamuge S. K. A trainable transparent universal approximator for defuzzification in Mamdani-type neuro-fuzzy controllers / S. K. Halgamuge // IEEE Transactions on Fuzzy Systems. – Vol. 6, № 2. – P. 304-314. DOI: 10.1109/91.669031.
58. Subbotin S. A. Data clustering based on inductive learning of neuro-fuzzy network with distance hashing / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2022. – № 4.

РОЗДІЛ 2
НЕЙРОЕВОЛЮЦІЙНІ МЕТОДИ ДЛЯ ОРГАНІЗАЦІЇ ПОШУКУ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ

Для ефективної експлуатації складних технологічних систем потрібні моніторинг і різні методи аналітики, що дозволяють контролювати, управляти, попереджувально змінювати параметри. Моніторинг, як правило, забезпечується типовими інструментами (в більшості випадків досить надійної системи збору і візуалізації даних). А ось для створення ефективних аналітичних інструментів необхідні додаткові дослідження, експерименти і добре знання предметної області. Як правило виділяють чотири основних види аналітики даних:

– описова аналітика візуалізує накопичені дані, в тому числі перетворені для наочності та інтерпретованості. Описова аналітика - найпростіший вид аналізу, але і найважливіший для застосування інших методів аналізу;

– за допомогою діагностичної аналітики досліджують причину виникнення подій в минулому, при цьому виявляються тренди, аномалії, найбільш характерні риси описуваного процесу, шукають причини і кореляції (взаємозв'язку);

– прогнозна аналітика передбачає ймовірні результати на основі виявлених тенденцій та статистичних моделей, отриманих за допомогою історичних даних;

– розпорядча аналітика дозволяє отримати оптимальне рішення виробничого завдання на основі прогнозної аналітики. Наприклад, це може бути оптимізація параметрів роботи обладнання або бізнес-процесів, перелік заходів щодо запобігання аварійної ситуації.

Для прогнозної і розпорядчої аналітики як правило використовуються методи моделювання, в тому числі Машинне навчання. Ефективність цих моделей залежить від якісної організації збору, обробки та попереднього аналізу даних. Перераховані види аналітики розрізняються як за складністю використовуваних моделей, так і за ступенем участі людини.

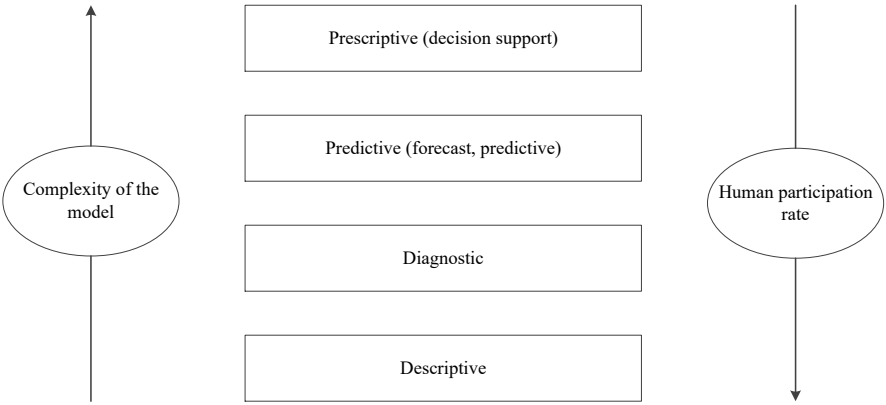


Рисунок 2.1 – Порівняння відмінностей моделей на основі участі людини і складності самих моделей

Сфер застосування інструментів аналітики дуже багато-інформаційна безпека, банківський сектор, Державне управління, медицина і багато інших предметні області. Часто один і той же метод ефективно працює для різних предметних областей, тому розробники систем аналітики створюють універсальні модулі, що містять різні алгоритми.

Для багатьох технологічних систем результати моніторингу можна представити у вигляді часових рядів. Властивостями тимчасового ряду є:

– прив'язка кожного виміру(семпла, дискрета) до часу його виникнення,

– рівна відстань за часом між вимірами,

– можливість з даних попереднього періоду відновити поведінку процесу в поточному і наступних періодах.

Часові ряди можуть описувати не тільки чисельно вимірні процеси. Застосування різних методів і архітектур моделей, включаючи глибокі нейронні мережі, дозволяє працювати з даними із завдань обробки природної мови (NLP), комп'ютерного зору (CV) і т.п. наприклад, повідомлення в чаті можна перетворити в числові вектори (ембеддинги), які послідовно з'являються в певний час, а відео представляє з себе ні що інше як матрицю чисел, що змінюється в часі.

Отже, часові ряди дуже корисні для опису роботи складних пристроїв і часто використовуються для типових завдань: моделювання, прогнозування, виділення ознак, Класифікація, кластеризація, пошук патернів, пошук аномалій. Приклади такого використання: електрокардіограма, зміна вартості акцій або валюти, значення прогнозу погоди, зміна в обсязі мережевого трафіку, параметри роботи двигуна і багато іншого.

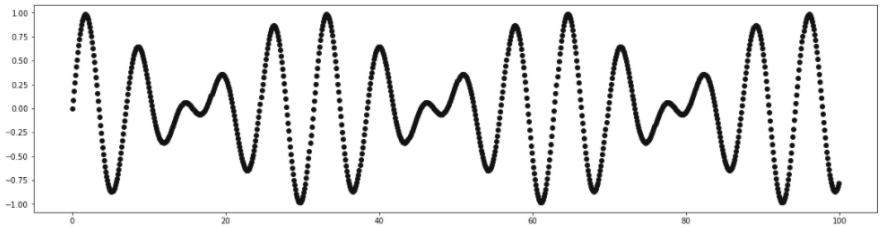


Рисунок 2.2 – Приклад часового ряду

У часових рядів є типові характеристики, які досить точно описують характер часового ряду:

– період – часовий відрізок постійної для всього ряду довжини, на кінцях якого ряд приймає близькі значення,

– сезонність – властивість періодичності (сезон=період),

– цикл – характерні зміни ряду, пов'язані з глобальними причинами (наприклад, цикли в економіці), немає постійного періоду,

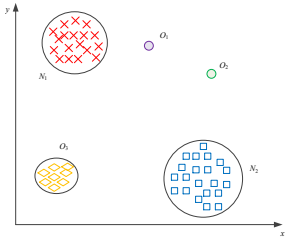
– тренд – тенденція до довгострокового збільшення або зменшення значень ряду.

Часові ряди можуть містити аномалії. Аномалія – це відхилення в стандартній поведінці якогось процесу. Метод машинного пошуку аномалій використовують дані про роботу процесу (датасети). Залежно від предметної області в датасеті можуть бути аномалії різного виду. Прийнято розрізняти кілька видів аномалій (рис. 2.3).

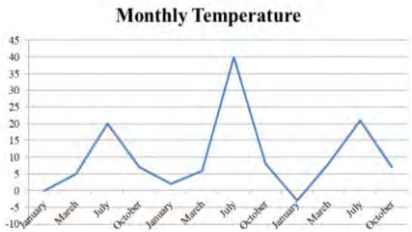
1. Точкові аномалії. Вони виникають у ситуаціях, коли окремий екземпляр даних може розглядатися як абсолютно ненормальний щодо інших [8].

2. Контекстуальні аномалії. Вони спостерігаються, якщо екземпляр є ненормальним у певному контексті або коли виконується певна умова (тому її також називають умовною) [8].

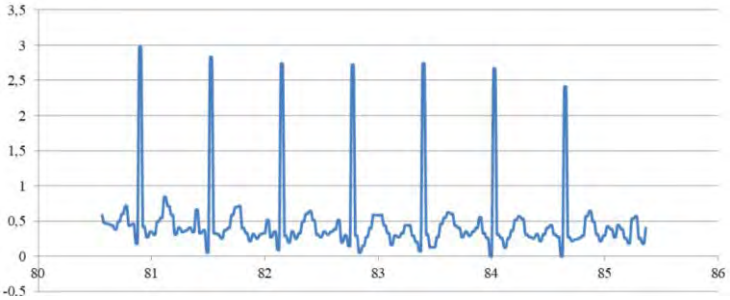
3. Колективні аномалії. Виникають, коли послідовність пов'язаних екземплярів даних (наприклад, графік часового ряду) є ненормальною по відношенню до решти даних [8].



а)



б)



в)

Рисунок 2.3 – Приклади різних аномалій у часових рядах: а) точкові та колективні аномалії у двовимірних даних; б) контекстуальна аномалія у часових рядах; в) колективна аномалія у часових рядах (ЕКГ): аритмія

Окремий екземпляр може і не бути відхиленням, але спільна поява таких екземплярів буде колективною аномалією.

2.1 Стратегії виявлення аномалій

Для виявлення точкових аномалій часто потрібна якась системна модель. Якщо система не має детермінованої математичної моделі або таку модель занадто складно побудувати, то повинна бути доступна статистична модель. Залежно від методу побудови статистичної моделі розрізняють наступні підходи [8-10].

1. Розпізнати аномалії під час навчання під наглядом. Цей метод вимагає наявності повноцінної навчальної вибірки, що включає

достатню кількість представників нормального і аномального класів значень.

Методика застосовується в 2 етапи: спочатку проводиться навчання на даних, які вручну вказують нормальні і ненормальні точки. Потім відбувається розпізнавання, коли нові дані класифікуються на основі побудованої моделі [8-10].

Зазвичай передбачається, що статистичні властивості моделі не змінюються з часом, і така зміна часто вимагає повторного навчання.

Основною складністю таких методів є формування даних для навчання. На додаток до очевидних трудових витрат, часто аномальний клас також представлений гірше, ніж нормальний, що може призвести до неточностей у результуючій моделі [8-10].

2. Розпізнавання аномалій частково контрольованого навчання. Подібно до попереднього, але навчальні дані представляють лише звичайний клас. Система, навчена в звичайному класі, може визначити, чи належать їй дані, таким чином ідентифікуючи аномальні дані шляхом виключення.

3. Розпізнавання у разі неконтрольованого навчання. При відсутності апріорної інформації це єдино можливий варіант. Розпізнавання при неконтрольованому навчанні: вільний режим заснований на припущенні, що аномальні дані зустрічаються досить рідко. Тому лише ті, які найбільш віддалені від середніх значень, позначаються як аномальні [8-10]. Застосування цього методу до потокових даних є складним, оскільки необхідно мати уявлення про весь масив даних, щоб мати хорошу оцінку середніх та очікуваних відхилень.

2.2 Методи розпізнавання аномалій

Класифікація. Цей метод заснований на тому, що нормальна поведінка системи може бути визначена одним або кількома класами. Тоді екземпляр, який не належить до жодного з класів, є аномальним. Цей метод зазвичай використовує підхід «частково контрольованого навчання». Основні механізми: нейронні мережі, байєсівські мережі, метод опорних векторів на основі правил [7, 9].

Кластеризація. Цей метод заснований на групуванні подібних значень в кластери і не вимагає знання властивостей можливих

відхилень. Виявлення аномалій базується на наступних припущеннях [8, 10]:

– звичайні екземпляри даних належать кластеру

– нормальні дані знаходяться ближче до центру кластера, ненормальні-далі

– нормальні дані утворюють великі щільні кластери, тоді як аномальні дані утворюють невеликі та розсіяні кластери.

– одним з найпростіших методів кластеризації є алгоритм *k*–середніх.

Статистичний аналіз. Використовуючи цей підхід, будується статистична модель процесу, яка потім порівнюється з фактичною поведінкою. Якщо фактична поведінка відрізняється від моделі більш ніж на певний поріг, робиться висновок про наявність аномалій. Методи статистичного аналізу поділяються на дві групи [7]:

– параметричні методи. Передбачається, що звичайні дані мають функцію $\rho(x,\theta)$ щільності ймовірності, де θ – вектор параметрів, x – екземпляр даних (спостереження);

– непараметричний метод. Структура моделі не визначена апріорі, але визначається на основі наданих даних.

Метод найближчого сусіда. При використанні цього методу вводиться метрика (міра подібності між об'єктами). Тоді можливі два підходи [10]:

– відстань до *k*-го найближчого сусіда. Аномальні дані найбільш далекі від усіх інших даних;

– використання відносної щільності. Проби в районах з низькою відносною щільністю оцінюються як аномальні. (наприклад, метод локального рівня викидів).

Спектральні методи. На основі спектральних (частотних) характеристик даних будується модель, яка призначена для обліку більшої частини мінливості даних.

Слід порівняти існуючі методи і представити результати у вигляді таблиці (табл. 2.1).

В цілому, з порівняння методів можна зробити висновок про недостатній якісний рівень існуючих методів, оскільки більшість методів не здатні використовувати всі стратегії (підходи) до машинного навчання, а також за результатами вони не дають однозначної відповіді про оцінку виявлених аномалій. Більш того, в

більшості робіт відзначається недостатній рівень точності при використанні класних методів з більш новими топологіями моделей. Саме тому завдання розробки нових підходів і методів виявлення аномалій у часових рядах залишається актуальним.

Таблиця 2.1 – Порівняння методів розпізнавання аномалій

Метод	Результат	Стратегія	Класифікації аномалій
Класифікація	Мітка (клас)	Навчання з вчителем, часткове навчання з вчителем	Так
Кластеризація	Мітка (кластер)	Навчання з вчителем, часткове навчання з вчителем	Ні
Статистичний аналіз	Ступінь	Часткове навчання з вчителем	Ні
Метод ближніх сусідів	Ступінь	Навчання без вчителя	Ні

2.3 Супутні роботи

Найпростіше розпізнати точкові аномалії-це окремі точки, в яких поведінка процесу різко відрізняється від інших точок. Наприклад, ви можете спостерігати різке відхилення значень параметрів у певній точці [7-10].

Такі значення називаються «викидами», вони сильно впливають на статистичні показники процесу і легко виявляються шляхом встановлення порогових значень для спостережуваного значення.

Складніше виявити аномалію в ситуації, коли процес поводиться нормально в кожній точці, але в сукупності значення в декількох точках поводяться дивно. Така ненормальна поведінка може включати, наприклад, зміну форми сигналу, зміну статистичних

показників (середнє значення, режим, медіана, дисперсія), появу взаємної кореляції між двома параметрами, невеликі або короткочасні аномальні зміни амплітуди і так далі. І в цьому випадку завдання полягає в розпізнаванні аномальної поведінки параметрів, які не можуть бути виявлені звичайними статистичними методами [7-10].

Пошук аномалій дуже важливий. В одній ситуації дані повинні бути очищені від аномалій, щоб отримати більш реалістичну картину, в той час як в іншій ситуації аномалії повинні бути ретельно вивчені, оскільки вони можуть вказувати на можливий швидкий перехід пристрою в аварійний режим.

Знайти аномалії в часових рядах непросто (нечітке визначення аномалій, відсутність розмітки, неочевидна кореляція). До цього часу алгоритм самоорганізованого дерева (Sota-алгоритми) для пошуку аномалій у часових рядах має високий рівень хибнопозитивних результатів [7-10].

Лише невелика кількість аномалій, переважно точкових, може бути виявлена вручну за наявності хорошої візуалізації даних. Групові аномалії важче виявити вручну, особливо якщо мова йде про великі обсяги даних та аналіз інформації про кілька пристроїв. Також важко виявити випадок аномалії в часі, коли нормальний сигнал з'являється в неправильний час. Тому при пошуку аномалій у часових рядах доцільно використовувати методи автоматизації [7-10].

Велика проблема пошуку аномалій у реальних даних полягає в тому, що дані, як правило, не позначаються, тому спочатку суворо не визначено, що таке аномалія, немає правил пошуку. У таких ситуаціях необхідно застосовувати методи навчання без участі викладача (непідтримуване навчання), в той час як моделі самостійно визначають взаємозв'язки і характерні закономірності в даних.

Методи, що використовуються для пошуку аномалій у часових рядах, зазвичай поділяються на групи [7-10]:

– на основі наближення: виявлення аномалій на основі інформації про параметри наближення або послідовності параметрів фіксованої довжини, підходить для виявлення точкових аномалій і викидів, але не буде виявляти зміни у формі сигналу;

– на основі прогнозування: побудуйте прогнозну модель та порівняйте прогнозоване та фактичне значення, яке найкраще застосовувати до часових рядів з яскраво вираженими періодами, циклами чи сезонністю;

– на основі реконструкції: методи, засновані на відновленні фрагментів даних, використовують відновлення фрагментів даних (реконструкцію), тому вони можуть виявляти як точкові аномалії, так і групові аномалії, включаючи зміни форми сигналу.

Методи, засновані на наближенні, орієнтовані на пошук значень, які значно відхиляються від поведінки всіх інших точок. Найпростішим і найбільш очевидним прикладом реалізації цього методу є моніторинг того, чи перевищено заданий поріг значень [8].

У методах, заснованих на прогнозуванні, основним завданням є побудова якісної моделі процесу для імітації сигналу і порівняння отриманих змодельованих значень з вихідними (істинними). Якщо передбачуваний і істинний сигнал близькі, то поведінка вважається нормальним, а якщо значення в моделі сильно відрізняються від істинних, то поведінка системи в цій області оголошується ненормальним [9].

Найпоширенішими методами моделювання часових рядів є SARIMA [11] та періодичні нейронні мережі.

Оригінальний підхід використовується в моделях, заснованих на реконструкції - спочатку модель навчають кодувати і декодувати сигнали з існуючої вибірки, при цьому закодований сигнал має набагато менший розмір, ніж вихідний, тому модель повинна навчитися стискати інформацію. Приклад такого стиснення для зображень розміром 32 на 32 пікселя наведено в [12].

Після навчання моделі видають вхідні сигнали, які є сегментами досліджуваного часового ряду, і якщо кодування і декодування проходять успішно, то поведінка процесу вважається "нормальним", в іншому випадку поведінка оголошується ненормальним.

Одним з нещодавно розроблених методів, заснованих на реконструкції, які показують хороші результати у виявленні аномалій, є TadGAN [13-15], розроблений дослідниками Массачусетського технологічного інституту наприкінці 2020 року. Архітектура методу TadGAN містить елементи автоматичного кодера і генеруючих змагальних мереж.

Мережа ε діє як кодер, який перетворює сегменти часового ряду x У приховані просторові вектори z, і є декодером, який відновлює сегменти часового ряду з прихованого подання z. є критиком ς , який оцінює якість відновлення $\varsigma(\varepsilon(x))$, і є критиком C_z , який оцінює

схожість прихованого подання $z = \varepsilon(x)$ з білим шумом. Крім того, існує контроль подібності вихідних та відновлених зразків за допомогою міри L2 відповідно до ідеології втрати послідовності циклу (що забезпечує загальну схожість генерованих зразків з вихідними зразками в GAN) [13]. Кінцева цільова функція являє собою комбінацію всіх показників для оцінки якості роботи критиків C_x, C_z, а також міру подібності між вихідним і відновленим сигналом.

$$\min_{\{\varepsilon,\varsigma\}} \max_{\{C_x \in \mathrm{C_x}, C_z \in \mathrm{C_z}\}} V_X(C_x,\varsigma) + V_Z(C_z,\varepsilon) + V_{L2}(\varepsilon,\varsigma)$$

Для створення та навчання нейронної мережі можна використовувати різні стандартні пакети (наприклад, TensorFlow або PyTorch), які мають API високого рівня. Приклад реалізації архітектури, подібної до TadGAN, за допомогою пакету силових тренувань TensorFlow можна знайти у сховищі [14, 15]. При навчанні цієї моделі були оптимізовані п'ять показників:

– aeLoss – це середньоквадратичне відхилення між початковим і відновленим часовими рядами, тобто різниця між x і $\varsigma(\varepsilon(x))$;

– cxLoss – бінарна крос-ентропійна критика C_x, яка визначає різницю між істинним сегментом часового ряду і штучно згенерованим,

– cx_g_loss – двійкова перехресна ентропія, помилка $\varsigma(\varepsilon(x))$ генератора, яка характеризує його нездатність" обдурити " критика ,

– czloss – двійкова перехресна ентропійна критика C_x, яка визначає різницю між прихованим вектором, що генерується кодером, і білим шумом, забезпечує схожість прихованого вектора з випадковий вектор, що перешкоджає запам'ятовуванню моделлю окремих патернів у вихідних даних,

– cz_g_Loss – це двійкова перехресна ентропія, помилка генератора $\varepsilon(x)$, яка характеризує його нездатність створювати приховані вектори, схожі на випадкові, і таким чином обманювати критика C_z.

Після навчання моделі реконструюються окремі сегменти досліджуваного часового ряду та порівнюються оригінальний та

реконструйований ряди, що може бути здійснено за допомогою одного з наступних методів [13]:

– порівняння потокової передачі;

– порівняння площ кривих в заданій області навколо кожного зразка (довжина області є гіперпараметром);

– динамічне спотворення часу.

Якість оцінюється з використанням показника $F1$ для задачі бінарної класифікації, позитивний (нульова гіпотеза): аномалія є, негативний (альтернативна гіпотеза): аномалії немає.

Таблиця 2.2 – Якість оцінюється з використанням показника $F1$ для задачі бінарної класифікації

	Аномалія передбачена моделлю	Модель передбачала відсутність аномалії
Аномалія присутня	TP правильно передбачена аномалія	FN є аномалія, але вона не була виявлена
Аномалія відсутня	FP передбачив аномалію там, де її не існує	TN аномалії немає, і модель її не бачить

Щоб продемонструвати роботу методу, ми використовуємо синтетичний (штучно сформований) ряд без аномалій, який є сумою двох синусоїд, значення яких коливаються від -1 до 1: $y(x)=\frac{1}{2}\sin(x)+\frac{1}{2}\sin(0.8x)$.

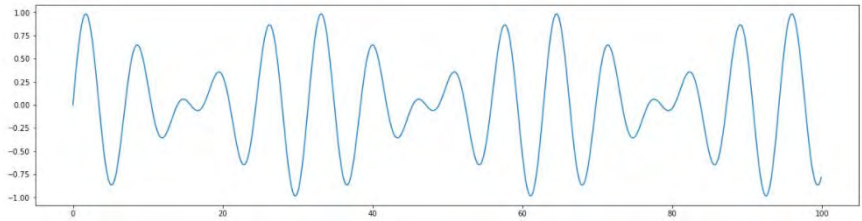


Рисунок 2.4 – Графік такого ряду

Видно, що модель досить точно навчилася передбачати основні закономірності в даних. Спробуємо додати різні аномалії до даних, а

потім виявити їх за допомогою моделі tadgan. Спочатку додамо кілька точкових аномалій [13].

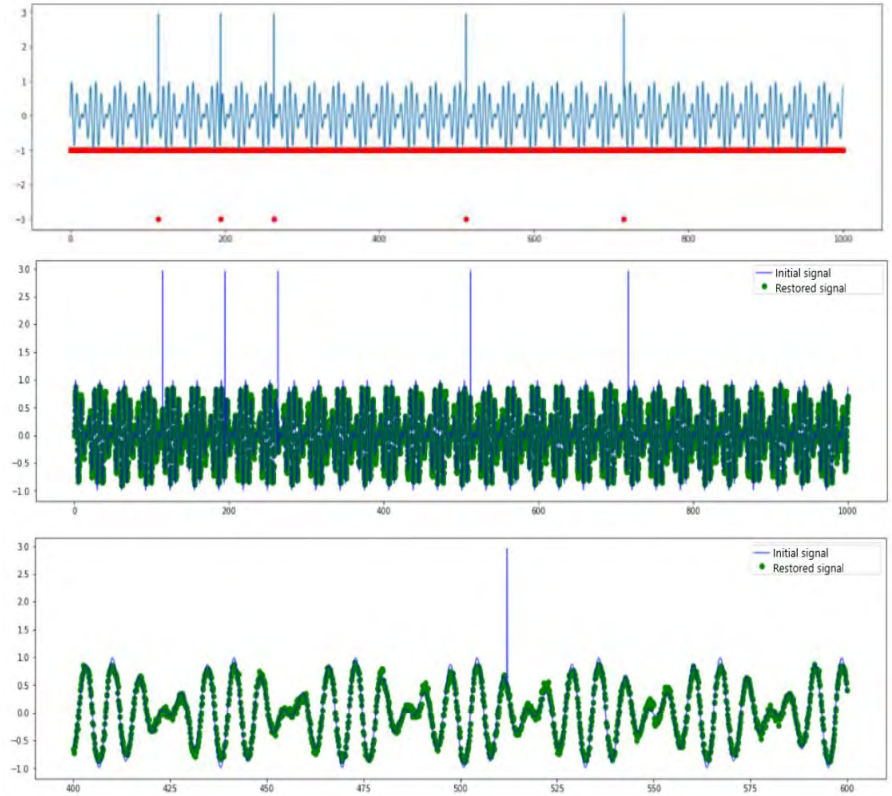


Рисунок 2.5 – Виявлення точкових аномалій за допомогою TadGAN

Графік вихідних і прогнозованих сигналів показує, що модель не може відновити "піки" аномальних значень, які можуть бути використані з високою точністю для визначення точкових аномалій. Однак у такій ситуації кора складної моделі Тадгана не очевидна-такі аномалії також можуть бути виявлені шляхом оцінки перевищення порогових значень [13].

Тепер розглянемо сигнал з іншим типом аномалії: періодичний сигнал з аномальною зміною частоти. У цьому випадку перевищення порогу немає: з точки зору амплітуди всі елементи ряду є

"нормальними" значеннями, і аномалія виявляється тільки в груповому поведінці декількох точок. У цьому випадку TadGAN також не може відновити сигнал (як видно на малюнку) і може бути використаний як ознака наявності групової аномалії [13].

Графік вихідних і прогнозованих сигналів показує, що модель не може відновити "піки" аномальних значень, які можуть бути використані з високою точністю для визначення точкових аномалій. Однак у такій ситуації кора складної моделі Тадгана не очевидна-такі аномалії також можуть бути виявлені шляхом оцінки перевищення порогових значень [13].

Тепер розглянемо сигнал з іншим типом аномалії: періодичний сигнал з аномальною зміною частоти. У цьому випадку перевищення порогу немає: з точки зору амплітуди всі елементи ряду є "нормальними" значеннями, і аномалія виявляється тільки в груповому поведінці декількох точок. У цьому випадку TadGAN також не може відновити сигнал (як видно на малюнку) і може бути використаний як ознака наявності групової аномалії [13].

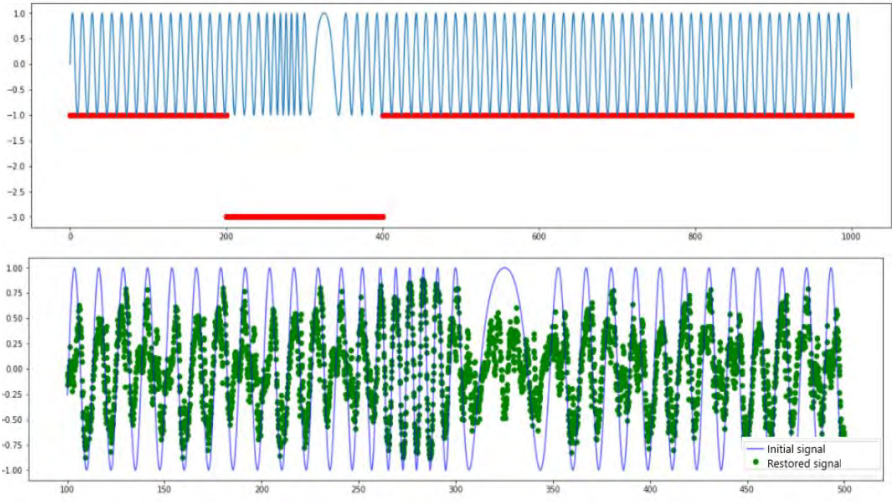


Рисунок 2.6 – Результат операції TadGAN з набором даних з аномальною зміною частоти.

Ці два приклади ілюструють роботу методу. Читач також може спробувати створити власні набори даних та протестувати можливості моделі в різних ситуаціях.

Більш складні приклади наборів даних можна знайти в статті авторів методу Тадгана [13]. Також є посилання на бібліотеку Orion, розроблену фахівцями Массачусетського технологічного інституту, яка використовує машинне навчання для розпізнавання рідкісних аномалій у часових рядах, використовуючи непідтримуваний підхід до навчання [14].

Однак більшість вчених сходяться на думці, що кожен конкретний випадок вимагає власного методу відновлення сигналу та провідного методу навчання моделі, що значно уповільнює практичну реалізацію.

2.4 Пропонований спосіб

Прототип рішення, що пропонується – показаний на рис. 2.7, складається з двох окремих груп однієї і тієї ж сукупності. Перший-це набір моделей, а другий - набір окремих даних.

Кожна модель являє собою послідовність рівнів кодера та декодера, які описують вхідний багатовимірний сигнал [16]. Фреймворк дозволяє створювати ансамблеву модель, використовуючи підхід, подібний до методу, заснованого на упаковці. Модель цілого ансамблю-це набір підмоделей, які працюють з підгрупами вхідних сигналів. Ансамблеві моделі утворюють набір.

Робота рішення починається з генерації вихідних груп вхідних сигналів за допомогою кореляції; потім система оновлює моделі та групи за допомогою генетичного алгоритму [17-19]. Паралельно генетичні оператори оптимізують окремі моделі в кожній підгрупі, вносячи зміни в топологію нейронних моделей (наприклад, довжину моделі та параметри шару) [17-19]. Кінцевим результатом цих дій є ансамблева модель, оптимізована для виявлення аномалій. Модель ансамблю визначається наступним чином.

У ряді робіт зазначається, що, незважаючи на різні моделі, помітно, що майже всі моделі демонструють схожий набір аномалій, незважаючи на зміни їх гіперпараметрів [16]. Звичайно, результати відрізнялися залежно від гіперпараметрів, але жодна зі змін суттєво не вплинула на виявлення. Через це пропонується ансамблева модель, заснована на одночасному поділі простору пошуку на підгрупи та еволюції моделей для таких підгруп. В результаті моделі змогли визначити більш конкретні залежності та взаємозв'язки між сигналами [20-22].

Моделі всередині кожної підгрупи оптимізуються шляхом зміни їх внутрішньої структури. Еволюція єдиної моделі здійснюється в п'ять основних етапів:

1. кластеризація простору пошуку [23-25];
2. схрещування, мутація та вибір найкращих моделей для окремих кластерів;
3. рішення для синхронізації;
4. кросинговер з кількома батьками [17-19];
5. оцінка ансамблевого рішення.

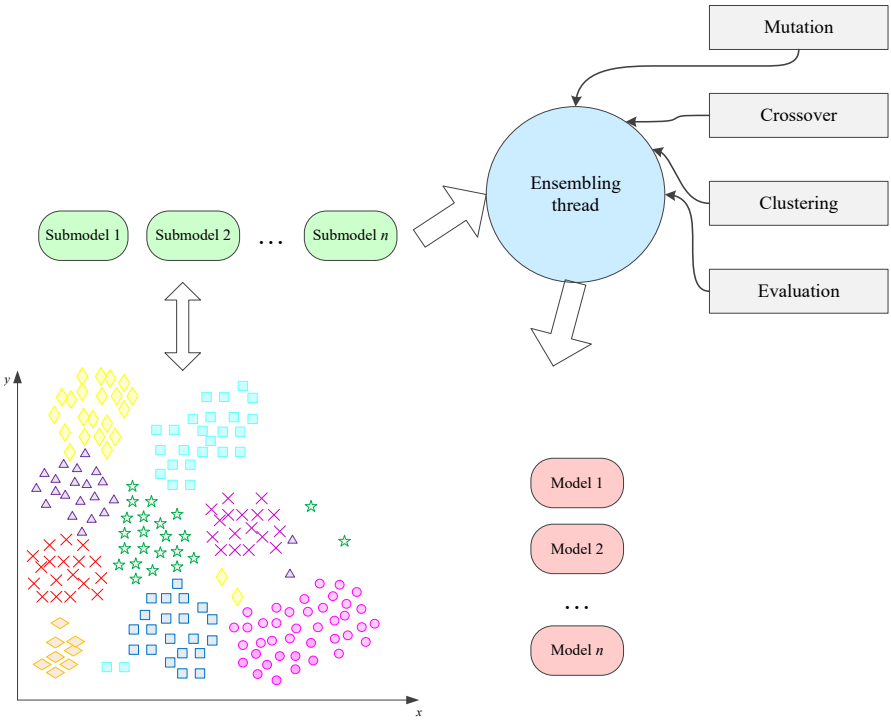


Рисунок 2.7 – Запропонований метод

2.5 Експериментальне дослідження методу

Для експериментального дослідження запропонованого методу в якості навчальних і тестових даних були використані наступні вибірки:

– набір даних з безпечного очищення води – The Secure Water Treatment (SWaT) [26], [16]. Цей набір даних містить дані, зібрані зі зменшеної версії реальної водоочисної установки. Дані були зібрані за 11 днів у двох режимах: 7 днів нормальної роботи заводу і 4 дні, протягом яких були здійснені кібератаки і сторонній фізичний вплив;

– набір даних про розподіл водних ресурсів – The Water Distribution (WADI) [27], [16]. Цей набір даних містить дані з меншої версії мережі розподілу води в місті. Зібрані дані містять 14 днів нормальної роботи та 2 дні, протягом яких було здійснено 15 атак.

Загальна інформація про набори даних представлена в табл. 2.3.

Таблиця 2.3 – Детальна характеристика вибірок

Вибірка	Кількість вхідних сигналів	Кількість тренувань	Кількість тестів	Кількість аномалій
SWAT	51	49,668	44,981	11.97%
WADI-2017	123	1,048,571	172,801	5.99%
WADI-2019	123	784,571	172,801	5.77%

Налаштовані метапараметри для нейроеволюційного синтезу наведено в табл. 2.4.

Результати тестування роботи методу наведено в табл. 2.5.

Таблиця 2.4 – Метапараметри синтезу

Метапараметр	Значення
Розмір популяції	100
Відсоток елітизму	5%
Функція активації	гіперболічний тангенс
Ймовірність мутації	25%
Тип схрещування	рівномірне
Типи мутації	видалення міжнейронного зв’язку
	видалення нейрону
	додавання міжнейронного зв’язку
	додавання нейрону
	зміна функції активації
Метод кластеризації	*k*-ближніх сусідів
Кількість сусідів	7

Таблиця 2.5 – Результати експериментального дослідження

Datasets	Точність, %	Повнота, %	f1, %
SWAT	94.41	55.35	0.74
WADI-2017	90.28	70.64	0.82
WADI-2019	89.53	71.47	0.83

2.6 Аналіз експериментальних результатів

Експериментальні результати демонструють, що паралельна кластеризація даних і синтез моделі на основі оброблених даних з використанням ансамблевої системи можуть значно підвищити ефективність процесу виявлення аномалій. Процес нейроеволюції допомагає синтезувати моделі та розробляти їх поетапно на основі оновленої інформації про вхідні дані, не розділяючи процес попередньої обробки даних. Тести на основі зразків даних WADI та SWAT. В обох випадках результати виявлення аномалій продемонстрували задовільні значення, що підвищило якісний рівень виявлення. Удосконалення набору даних WADI були більш значними, ніж у наборі даних SWAT, оскільки набір даних WADI має більше датчиків і зразків, ніж набір даних SWAT.

Під час роботи доведено, що нейроеволюційний підхід може мати позитивний вплив на результати. Наша майбутня робота буде зосереджена на подальшому вдосконаленні цього методу. Поєднання різних топологій рішень може значно поліпшити результати роботи.

2.7 Висновки за розділом 2

В розділі було проведено дослідження та порівняльний аналіз існуючих стратегій та методів, що вирішують проблему виявлення та класифікації аномалій, а також запропоновано метод виявлення, заснований на нейроеволюційних методах. Як видно з результатів дослідження, метод виявлення аномалій, заснований на нейроеволюції, показав набагато більшу ефективність. Запропонований метод показав себе життєздатним і може бути вдосконалений.

Результатом роботи є не тільки всебічне вивчення і теоретичне обґрунтування теорії, пов'язаної з аналізом часових рядів, а й запропоноване рішення. Ця робота складається з двох етапів: розділення простору пошуку та синтезу моделей. На етапі навчання метод обробляє і розділяє дані про поведінку системи. У режимі синтезу метод поступово коригує моделі, щоб в майбутньому отримати з них остаточне рішення. Отримана модель синтезується з використанням рівномірного схрещування, що дозволяє збільшити розмір батьківського пулу з двох особин до набагато більшого числа.

2.8 Література до розділу 2

1. Chandola V. Anomaly detection: A survey / V. Chandola, A. Banerjee, V. Kumar // ACM Computing Surveys. – 2009. – Vol. 41(3). – P. 1-58. DOI: 10.1145/1541880.1541882.
2. What Is Data Analytics? [Electronic resource]. – Access mode: https://www.intel.com/content/www/us/en/artificial-intelligence/what-is-data-analytics.html#:~:text=Data%20analytics%20is%20the%20process,data%20for%20practically%20any%20purpose
3. Cycle Consistency Loss [Electronic resource]. – Access mode: https://paperswithcode.com/method/cycle-consistency-loss
4. Search for anomalies in time series based on the estimation of their parameters [Electronic resource]. – Access mode: https://openarchive.nure.ua/items/7c8e2e76-10fa-4044-907b-e51d05bd7cbf
5. Ma J. Time-series novelty detection using one-class support vector machines [Electronic resource] / J. Ma, S. Perkins // 2003 International Joint Conference on Neural Networks, Portland, OR, USA. – Portland: IEEE, 2003. – Vol. 3. – P. 1741-1745. DOI: 10.1109/ijcnn.2003.1223670
6. MIT – Data to AI Lab, Time series anomaly detection – in the era of deep learning [Electronic resource]. – Access mode: https://medium.com/mit-data-to-ai-lab/time-series-anomaly-detection-in-the-era-of-deep-learning-dccb2fb58fd
7. Anomaly Detection in Time Series [Electronic resource]. – Access mode: https://neptune.ai/blog/anomaly-detection-in-time-series

8. Bhattacharya A. Effective Approaches for Time Series Anomaly Detection / A. Bhattacharya : [Electronic resource]. – Access mode: https://towardsdatascience.com/effective-approaches-for-time-series-anomaly-detection-9485b40077f1
9. Schmidl S. Anomaly detection in time series / S. Schmidl, P. Wenig, T. Papenbrock // Proceedings of the VLDB Endowment. – 2022. – Vol. 15, no. 9. – P. 1779–1797. DOI: 10.14778/3538598.3538602
10. Anomaly detection and forecasting in Azure Data Explorer [Electronic resource]. – Access mode: https://learn.microsoft.com/en-us/azure/data-explorer/kusto/query/anomaly-detection
11. Artley B. Time Series Forecasting with ARIMA , SARIMA and SARIMAX / B. Artley : [Electronic resource]. – Access mode: https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6
12. Mandal M.K. Implementing PCA, Feedforward and Convolutional Autoencoders and using it for Image Reconstruction, Retrieval & Compression / M. K. Mandal : [Electronic resource]. – Access mode: https://blog.manash.io/implementing-pca-feedforward-and-convolutional-autoencoders-and-using-it-for-image-reconstruction-8ee44198ea55
13. TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks / [Alexander Geiger et al.] // 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020. DOI: 10.1109/bigdata50022.2020.9378139
14. TadGAN [Electronic resource]. – Access mode: https://github.com/gusty1g/TadGAN
15. Examples. TadGAN [Electronic resource]. – Access mode: https://github.com/CyberLympha/Examples/tree/main/%D0%A0%D0%B0%D0%B7%D0%B1%D0%BE%D1%80%20%D1%81%D1%82%D0%B0%D1%82%D0%B5%D0%B9/TadGAN
16. Faber K. Ensemble Neuroevolution-Based Approach for Multivariate Time Series Anomaly Detection / Kamil Faber, Marcin Pietron, Dominik Zurek // Entropy. – 2021. – Vol. 23, no. 11. – P. 1466. DOI: 10.3390/e23111466
17. Smart Crossover Mechanism for Parallel Neuroevolution Method of Medical Diagnostic Models Synthesis / [S. Leoshchenko, S. Subbotin, A. Oliinyk, V. Lytvyn, M. Ilyashenko] // Proceedings of the Third International Workshop on Computer Modeling and Intelligent

Systems (CMIS-2020), Zaporizhzhia, Ukraine, April 27-May 1, 2020. – P. 57-69.
18. Leoshchenko, S. Adaptive Mechanisms for Parallelization of the Genetic Method of Neural Network Synthesis / S. Leoshchenko, A. Oliinyk, S. Subbotin // Proceedings of the 10th International Conference on Advanced Computer Information Technologies (ACIT 2020), Deggendorf, Germany, 16-18 September 2020 : proceedings of the conference. – Ternopil : West Ukrainian National University, 2020. – P. 446-450.
19. Sequencing for Encoding in Neuroevolutionary Synthesis of Neural Network Models for Medical Diagnosis / [S. Leoshchenko, A. Oliinyk, S. Subbotin, T. Zaiko, S. Shylo, V. Lytvyn] // Proceedings of the 3rd International Conference on Informatics & Data-Driven Medicine (IDDM 2020), Växjö, Sweden, 19-21 October 2020. – P. 62-71.
20. Alsayaydeh J.A.J. Intelligent Interfaces for Assisting Blind People using Object Recognition Methods / [J.A.J. Alsayaydeh, Irianto, M. Zainon, H. Baskaran, S. G. Herawan] // International Journal of Advanced Computer Science and Applications. – 2022. – Vol. 13, no. 5. – P. 734-741. DOI: 10.14569/IJACSA.2022.0130584
21. Alsayaydeh J.A.J. Face Recognition System Design and Implementation using Neural Networks / [J.A.J. Alsayaydeh, Irianto, A. Aziz, C.K. Xin, A. K. M. Zakir Hossain, S.G. Herawan] // International Journal of Advanced Computer Science and Applications. – 2022. – Vol. 13, no. 6. – P. 519-526. DOI: 10.14569/IJACSA.2022.0130663
22. Shkarupylo V. Iterative Approach to TLC Model Checker Application / [V. Shkarupylo, I. Blinov, A. Chemeris, J.A.J. Alsayaydeh, A. Oliinyk] // 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 13–17 September 2021. – P. 283-287. DOI: 10.1109/KhPIWeek53812.2021.9570055R.
23. k-NN based Time Series Classification [Electronic resource]. – Access mode:https://towardsdatascience.com/k-nn-based-time-series-classification-e5d761d01ea2
24. Applying k-nearest neighbors to time series forecasting : two new approaches / [S. Tajmouati, B. Wahbi, A. Bedoui, A. Abarda, M. Dakkon] : [Electronic resource]. – Access mode: https://arxiv.org/abs/2103.14200

25. A methodology for applying k-nearest neighbor to time series forecasting / [F. Martínez, M.P. Frías, M. Pérez, A. J. Rivera Rivas] // Artificial Intelligence Review. – 2017. – Vol. 52, no. 3. – P. 2019–2037. DOI: 10.1007/s10462-017-9593-z
26. Mathur A. P. SWaT: a water treatment testbed for research and training on ICS security / A. P. Mathur, N. O. Tippenhauer // 2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater), Vienna, Austria, 11 April 2016. – [S. l.], 2016. – DOI: 10.1109/cyswater.2016.7469060
27. Ahmed C. M. WADI / C. M. Ahmed, V. R, Palleti, A. P. Mathur // CPS Week '17: Cyber Physical Systems Week 2017, Pittsburgh Pennsylvania. – New York, NY, USA, 2017. – DOI: 10.1145/3055366.3055375

РОЗДІЛ 3 ВИКОРИСТАННЯ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ ІНФОРМАЦІЙНО-ОРІЄНТОВНИХ ЗАСТОСУНКІВ

Незалежно від того, скільки разів конструюється образ покупця, сегментація є дуже усередненою. Людський мозок не може обробляти величезні обсяги даних, робити мільйони висновків і сценаріїв, запам'ятовувати їх і ефективно застосовувати. Але це може бути зроблено машинами, шляхом їх навчання та створення штучних нейронних мереж (ШНМ) для інформаційно-орієнтовних застосунків в бізнесі.

ШНМ – це один із способів сприйняття сенсорної інформації штучним або машинним інтелектом. Очевидно, що біологічні нейронні мережі стали прототипом нейронних мереж. Тобто людські способи отримання візуальної інформації, яка становить дві третини від загального сенсорного трафіку [1].

ШНМ – це підхід до вирішення завдань, які не можуть бути розв’язані за допомогою класичного алгоритмічного програмування із високою точністю. Немає необхідності будувати моделі для всіх можливих випадків розвитку системи або процесів в системі і прагнути передбачити всі варіанти і описати логіку для них [2-6]. ШНМ дозволяють на основі великої кількості накопичених даних самостійно знаходити закономірності і взаємозв'язки в раніше не явних аспектах і використовувати цю інформацію для подальшого прогнозування, класифікації та управління даними і процесами.

Примітивно, можна приблизно описати процес роботи ШНМ наступним чином: нейронна мережа на вхід отримує великі дані. Потім ці дані аналізуються, нейронна мережа навчається на основі позитивних і негативних прикладів. У процесі навчання формується структура нейронної мережі, яка в майбутньому може вирішувати завдання ідентифікації, класифікації, прогнозування [7].

3.1 Перехід до бізнесу, орієнтованого на дані

ШНМ буде допомагати підприємствам просувати товари та послуги. Якщо маркетологи зараз покладаються на середню

сегментацію і таргетинг, то в найближчому майбутньому ШНМ, знаючи так багато про користувача і вміючи дуже швидко обробляти інформацію, будуть точно передбачати, чого хоче людина. Глибоке розуміння бажань і потреб споживача є ключем до успіху. Здатність бачити проблеми, поведінку та життєвий цикл у динаміці відкриває великі можливості для бізнесу [8-10].

Саме так вже працюють деякі сервіси, такі як Google Play Music [11]. Нейронна мережа вивчає користувача день за днем і, нарешті, дає рекомендації, які точно відповідають інтересам. ШНМ на цьому не зупиняється: їхні рекомендації передбачають інтереси. Людина приходить послухати одну групу, але в, врешті-решт, із задоволенням повертається до пісень, які слухала колись, і які викликають великий напад ностальгії.

ШНМ допоможуть компаніям уникнути комерціалізації технологій. Це ситуація, коли всі товари, послуги та товари швидко стають однаковими. Зовнішній вигляд, якість і вигода — все те ж, що і у конкурентів. Щоб залучити клієнтів допоможе особливий підхід до кожного. І це дозволить створити нейронну мережу [12], [13].

Наприклад, клієнт купує новий телевізор. Співробітник магазину за звичкою рекомендує придбати витратні матеріали: пульт дистанційного керування, кабель, систему кріплення. Система на основі ШНМ в лічені секунди проаналізує клієнта і його покупку, і запропонує придбати кавоварку. ШНМ визначить, що клієнт просто обставляє кухню і думає про покупку хорошої кавомашини [14], [15].

Визначимо ситуації, коли доцільно використовувати ШНМ у бізнес-завданнях [16]:

– накопичено величезну кількість різних даних;

– поки немає робочих методів обробки і систематизації цих даних;

– дані пошкоджені, зіпсовані, неповні або несистематизовані;

– існує велика різноманітність даних, і на перший погляд важко встановити зв'язки та закономірності між ними.

Можливості та приклади можливого застосування нейронних мереж та машинного навчання для бізнес-завдань [17]:

– прогнозування, оцінка ризиків (прогнозування попиту, обсягу продажів, середнього чека, частоти продажів, завантаження обладнання для оптимізації кількості готівки, складських приміщень та інших ресурсів);

– пошук тенденцій, кореляцій, трендів. Прогнозування подальшого розвитку системи та прогнозування можливих змін. Штучний інтелект значно покращив механізми рекомендацій в інтернет-магазинах та сервісах. Методи, засновані на машинному навчанні, аналізують поведінку клієнтів на сайті та порівнюють її з мільйонами інших користувачів. Все це для того, щоб визначити, який продукт користувач буде купувати з найбільшою ймовірністю:

– розпізнавання фото-, відео-, аудіоконтенту. Різні сервіси та онлайн-застосунки, що використовують технологію розпізнавання;

– методи машинного навчання для ведення діалогу комп'ютерними системами. Для автоматизації діяльності операторів в онлайн-чатах і месенджерах, телефонних операторів. Розробка чат-ботів. Системи, що аналізують природну мову, можуть бути використані для створення чат-ботів, які дозволяють клієнтам отримувати необхідну інформацію про продукти компанії. Це дозволить знизити витрати на команду колл-центру [18].

3.2 Рекурентні ШНМ для інформаційно-орієнтованих застосунків та бізнесу

Рекурентні ШНМ (РНМ) – це тип ШНМ, призначений для розпізнавання конкретних моделей у послідовності даних, незалежно від їх виду: текст, зображення, аудіопотоки (включаючи вимовлені слова) або ж числові послідовності, що надходять від датчиків. Основною відмінністю РНМ від інших архітектур є так звана наявність пам'яті. РНМ зберігає попереднє значення у своєму стані. Для наочності розглянемо приклад [19], [20].

У процесі життя люди не починають щомиті мислити з нуля. Тобто стирання всіх раніше накопичених знань не відбувається, і будь-яка розумова діяльність грунтується на існуючих знаннях і досвіді. Всі наші знання і думки постійні.

Традиційні ШНМ не володіють цією властивістю, і це їх головний недолік. РНМ, навпаки, зберігають колишні значення, які в розглянутому прикладі можна прирівняти до аналогу людських знань і досвіду [19], [20].

Можна сказати, що РНМ будує динамічні моделі, тобто моделі, які змінюються з часом таким чином, що можна досягти достатньої точності, залежно від контексту наведених прикладів.

Розглянемо фрагмент РНМ (рис. 3.1). Фрагмент нейронної мережі А приймає вхідне значення x_t і повертає значення h_t. Наявність зворотного зв'язку дозволяє передавати інформацію від одного кроку мережі до іншого.

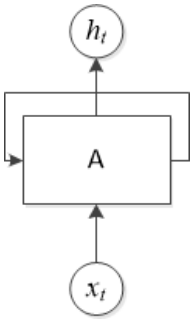


Рисунок 3.1 – Фрагмент РНМ

Складність РНМ полягає в наявності зворотних зв'язків [21], [22]. Тому, для більшої наочності, можна представити РНМ у вигляді окремих копій однієї і тієї ж мережі, кожна з яких передає інформацію в наступну копію [21], [22]. Схему з розширеними зворотними зв'язками наведено на рис. 3.2.

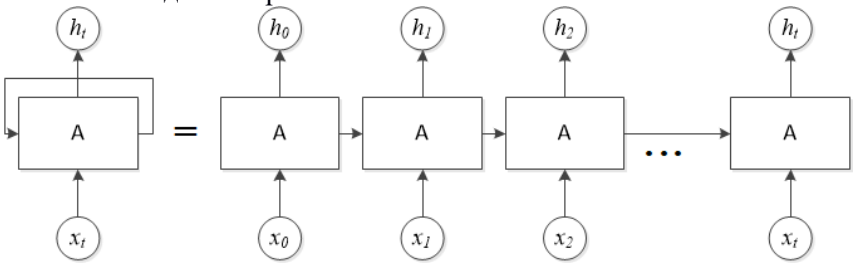


Рисунок 3.2 – Схема з розширеними зворотними зв'язками

Однак під час використання РНМ часто стикаються з проблемою довготривалих залежностей [21], [22].

Для виконання більшості поточних завдань потрібна тільки свіжа інформація. У цьому випадку можна стверджувати, що відстань між актуальною інформацією і місцем, де вона потрібна, невелика.

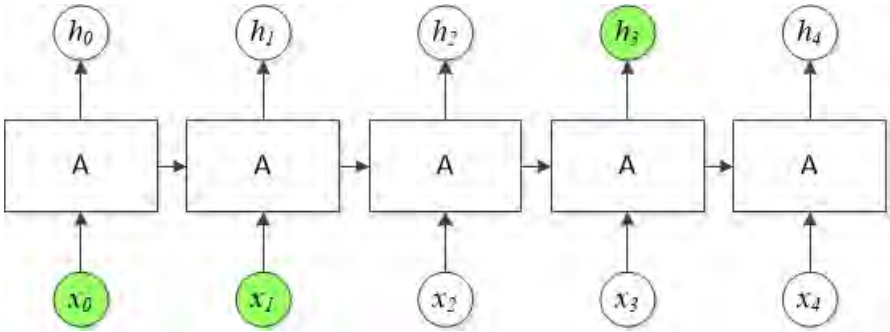


Рисунок 3.3 – РНМ може навчатися використанню інформації з минулого

Але в реальній практиці такі випадки рідкісні. В експериментах з розпізнавання мови було відзначено, що для передбачення закінчення фраз необхідний великий контекст. І в таких випадках розрив між фактичною інформацією і точкою її застосування може стати дуже великим, і з ростом цієї відстані РНМ втрачає здатність зв'язувати інформацію.

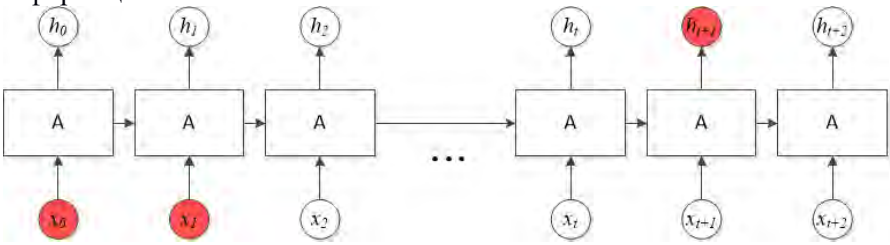


Рисунок 3.4 – РНМ втрачають здатність передавати інформацію

Теоретично проблеми з обробкою довгострокових залежностей в РНМ можуть бути вирішені шляхом ретельного відбору або штучної підміни розв'язуваної задачі. На жаль, на практиці це занадто ресурсоємно. Ця проблема детально досліджена Сеппом Хохрайтером [23] та Йошуа Бенгіо зі співавторами [23]; вони знайшли незаперечні причини, чому це може бути важко.

На сьогодні РНМ з неймовірним успіхом використовуються для вирішення цілого ряду завдань: розпізнавання мови, мовного моделювання, перекладу, розпізнавання зображень і т. д.

Значна роль в цих успіхах належить мережам Long short-term memory (LSTM) – незвичайній модифікації РНМ, яка в багатьох

завданнях значно перевершує стандартну версію. Майже всі вражаючі результати РНМ досягаються за допомогою LSTM. LSTM розроблені спеціально для того, щоб уникнути проблеми довгострокової залежності. Зберігання інформації протягом тривалого періоду часу – це їх нормальна поведінка, а не те, чого вони намагаються навчитися.

3.3 Топології РНМ для інформаційно-орієнтовних застосунків

3.3.1 LSTM

LSTM – це особливий вид архітектури РНМ, здатний вивчати довгострокові залежності. Представлений С. Хохрайтером та Ю. Шмідхубером у 1997 році [24], а також вдосконалений та адаптований, популярний у роботі багатьох інших дослідників. Вони прекрасно вирішують цілий ряд різних завдань і в даний час широко використовуються.

У порівнянні зі звичайною принциповою схемою мереж РНМ (рис. 3.1) і LSTM (рис. 3.5) можна відзначити більш складне представлення LSTM через наявність додаткових елементів, які називаються гейтами (шлюзи), які необхідні для управління потоками даних. Залежно від свого стану гейт може пропустити сигнал, а може і не пропустити його.

LSTM складається з наступних частин (рис. 3.5).

– мережевий вхід (input);
– мережевий вихід (output);
– стан пам'яті або мережі (memory cell);
– гйет для очищення пам'яті (forget gate);
– гейт оновлення пам'яті (input gate);
– вихідний гейт (output gate).

Як видно з опису мережі, LSTM дійсно краще справляється з поставленими завданнями завдяки складній архітектурі. Однак така архітектура вимагає додаткових обчислень і, як наслідок, додаткових ресурсів для обчислень і зберігання даних. Більш того, дослідження показали, що перед навчанням мереж LSTM дані повинні бути відфільтровані, тобто повинен бути проведений відбір інформативних ознак, а це також вимагає додаткових обчислювальних потужностей.

У 2014 році К.Чо представив модель Gated Recurrent Unit (GRU), засновану на тих же принципах, що і LSTM, але використовує менше фільтрів і операцій для обчислень [24].

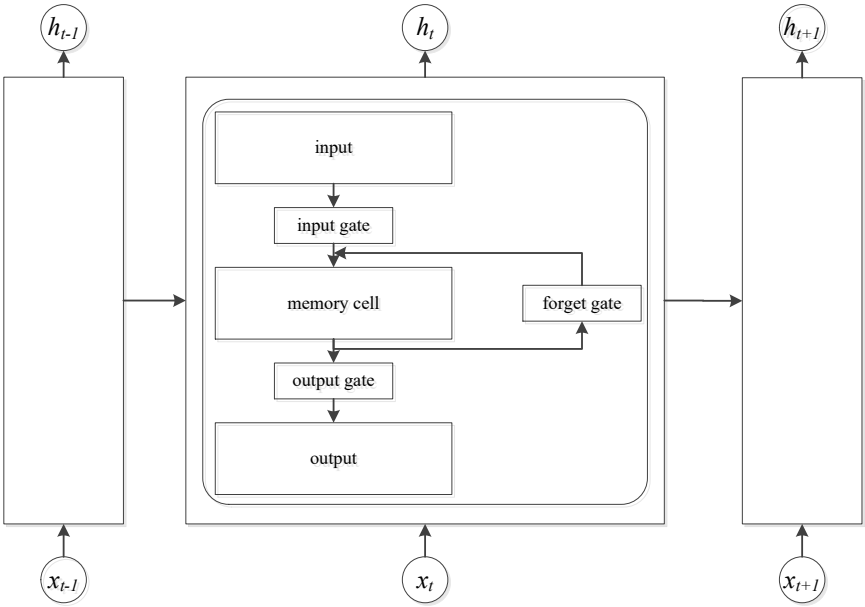


Рисунок 3.5 – Фрагмент мережі LSTM

3.3.2 GRU

GRU – це невелика варіація попередньої мережі. Тут на один фільтр менше, і зв'язки реалізовані по-іншому. Гейт оновлення (update gate) визначає, скільки інформації залишиться від попереднього стану і скільки буде взято з попереднього рівня. Гейт оновлення (reset gate) працює як гейт забуття. Схематичне зображення мереж обох типів демонструє відмінності (рис. 3.6) [25].

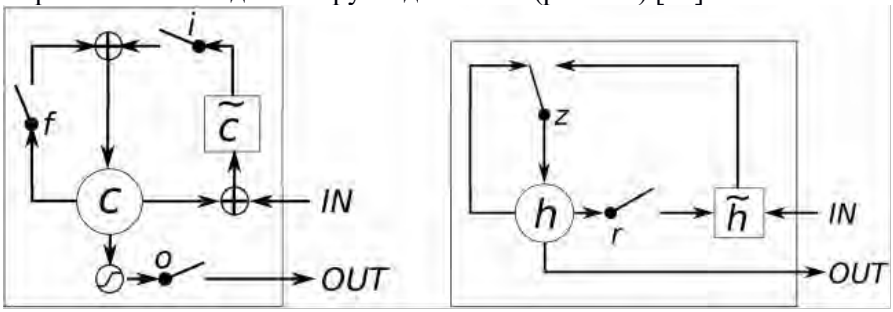


Рисунок 3.6 – Порівняння LSTM та GRU

У разі порівняння LSTM та GRU перевага надається GRU у випадках:

– коли необхідно прискорити час роботи;

– коли неможливо вибрати інформативні ознаки або їх недостатньо;

– коли ресурси пам'яті обмежені;

– у завданнях розпізнавання мови.

3.3.3 Двонаправлені РНМ

Мережі типу Bidirectional recurrent neural networks (BRNN) базуються на ідеї, що вихідні дані в момент часу t можуть залежати не тільки від попередніх елементів послідовності, але і від майбутніх. Наприклад, якщо потрібно передбачити пропущене слово в послідовності, враховуючи як лівий, так і правий контекст. З виходами все досить просто. Це просто дві РНМ, що накладені одна на іншу. Потім вихідні дані обчислюються на основі прихованого стану обох запусків [26].

Різниця полягає в тому, що ці мережі використовують не тільки дані з минулого, але і з майбутнього. Такі мережі здатні, наприклад, не тільки розширювати зображення по краях, але і заповнювати отвори всередині [26].

3.4 Використання РНМ для інформаційно-орієнтованих застосунків та бізнесу

Стан будь-якого елемента бізнесу, орієнтованого на дані, характеризується великою кількістю параметрів стану (ознак), значення яких можуть бути отримані за результатами спостережень за клієнтами або ринком [27]. ШНМ дозволяють класифікувати і прогнозувати подальші тенденції, нові бажання і потреби клієнтів, глобальні зміни на ринку.

Після навчання мережа здатна передбачити майбутнє значення послідовності на основі кількох попередніх значень або деяких існуючих факторів. Слід зазначити, що прогнозування можливе лише тоді, коли попередні зміни певною мірою визначають майбутнє.

У ШНМ проблема прогнозування формалізується через проблему розпізнавання шаблонів. Дані про прогнозованої змінної за певний

період часу формують зображення, клас якого визначається значенням прогнозованої змінної в деякий момент часу за межами даного періоду, тобто значенням змінної через інтервал прогнозування.

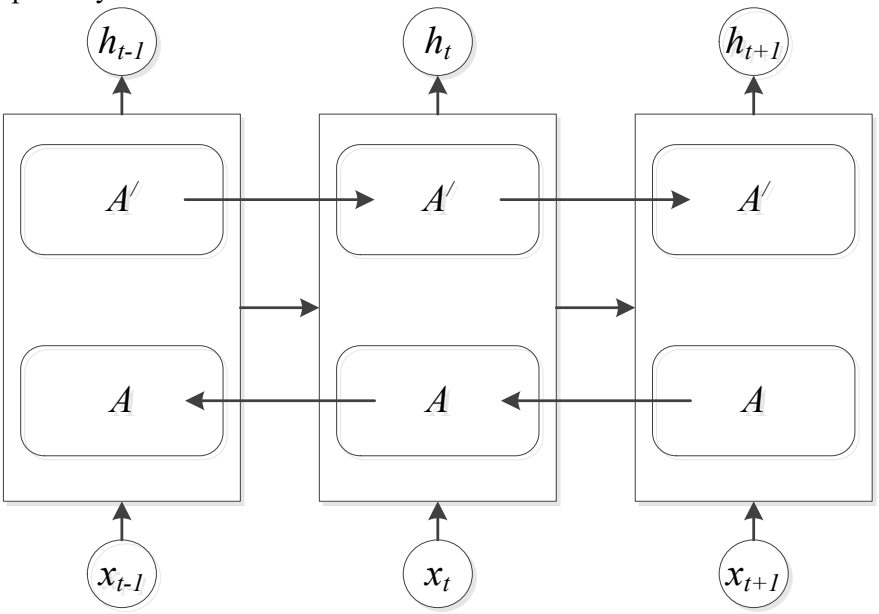


Рисунок 3.7 – Огляд структури BRNN

3.5 Експериментальне дослідження використання різних топологій РНМ

В якості моделі (інструменту побудови) передбачається використовувати мережі LSTM, GRU і BRNN. Тест дозволить визначити потенціал кожної архітектури. Для експерименту планується використовувати два тестових зразка, які знаходяться у вільному доступі [28]. Один з тестових зразків пройшов процес вибору функцій [28], спеціально для порівняння роботи двох архітектур з різними даними. Як метод навчання нейронних мереж використовувалося зворотне поширення в часі [29]. Інформація про обидва тестованих зразках наведена в табл. 4.1.

Таблиця 4.1 – Характеристики вибірок даних

Dow Jones Index Data Set			
Критерій	Характеристика	Критерій	Характеристика
Характеристика вибірки	Multivariate	Кількість екземплярів	750
Характеристика атрибутів	Integer, Real	Кількість атрибутів у екземплярах	16
Online Retail Data Set			
Критерій	Характеристика	Критерій	Характеристика
Характеристика вибірки	Multivariate	Кількість екземплярів	541909
Характеристика атрибутів	Integer, Real	Кількість атрибутів у екземплярах	8

В ході тестування особливу увагу було приділено основним параметрам роботи: витраченого часу, середнього значення похибки і пороговим значенням похибки.

Таблиця 4.2 – Результати експериментів

Dow Jones Index Data Set				
Критерії	Час синтезу	E_{min}	E_{max}	E
LSTM	10634.08	0.006	0.011	0.009
GRU	5139.69	0.018	0.026	0.022
BRNN	19125.30	0.003	0.009	0.006
Online Retail Data Set				
Критерії	Час синтезу	E_{min}	E_{max}	E
LSTM	24525.65	0.009	0.015	0.012
GRU	12354.36	0.014	0.032	0.023
BRNN	30262.52	0.008	0.013	0.011

3.6 Аналіз отриманого результатів

З отриманих результатів можна зробити наступні висновки. По-перше, в обох випадках час, витрачений на LSTM та BRNN, довший, ніж на GRU. Однак у випадку LSTM та BRNN збільшується пропорційно збільшенню обсягу вхідних даних. Для GRU зміна часу досить різка.

По-друге, слід зазначити, що розмір середньої помилки для мереж LSTM або BRNN в обох випадках менший, що характеризує їх роботу як більш точну. При роботі з технічною діагностикою складних систем, від яких може залежати життя людини, цей параметр є найбільш важливим. Крім того, важливо відзначити, що при тестуванні архітектури GRU з вхідними даними, які не були оброблені, розрив між мінімальним значенням помилки і максимальним занадто великий.

По-третє, при роботі з ШНМ проблема перенавчання є досить поширеною. Зі збільшенням числа інформаційних ознак точність прогнозування часто знижується. Особливо, якщо в даних багато неінформативних ознак (мало пов'язаних з цільовою змінною). Отже, процес вибору ознак-це етап в роботі з ШНМ, який ні в якому разі не можна упускати, особливо з огляду на недавні вражаючі результати застосування нових методів і підходів [30]. Таким чином, використання попередньо оброблених зразків як вхідних даних значно скоротить час, необхідний для подальшої обробки.

Особливу увагу слід приділити порівнянню результатів LSTM і BRNN. Під час тестування було відзначено, що результати точності LSTM поступаються точності BRN. Однак час навчання LSTM набагато коротший. З цього випливає, що найбільш оптимальним було б використання мереж LSTM.

3.7 Висновки за розділом 3

ШНМ незамінні для аналізу даних, зокрема для попереднього аналізу або відбору, для виявлення викидів або грубих помилок особи, яка приймає рішення. Доцільно використовувати нейромережеві методи в задачах з неповними або неінформативними даними, особливо в задачах, де рішення може бути знайдено

інтуїтивно, а традиційні математичні моделі не дають бажаного результату.

Методи нейронних мереж можуть бути використані самостійно або служити чудовим доповненням до традиційних методів статистичного аналізу, більшість з яких передбачає побудову моделей на основі певних припущень та теоретичних висновків (наприклад, що бажана залежність є лінійною або що якась змінна має нормальний розподіл). Нейромережевий підхід не пов'язаний з такими припущеннями - він однаково підходить як для лінійних, так і для складних нелінійних взаємозв'язків, але особливо ефективний при дослідницькому аналізі даних, коли метою є з'ясувати, чи існують залежності між змінними. У цьому випадку дані можуть бути неповними, суперечливими і навіть навмисно спотвореними. Якщо існує якийсь зв'язок між вхідними та вихідними даними, навіть якщо вони не виявляються традиційними методами кореляції, нейронна мережа здатна автоматично налаштуватися на неї із заданим ступенем точності. Крім того, сучасні нейронні мережі володіють додатковими можливостями: вони дозволяють оцінювати порівняльну важливість різних типів вхідної інформації, зменшувати її обсяг без втрати значущих даних, розпізнавати симптоми наближення критичних ситуацій і т. д.

Результати досліджень і їх аналіз показують, що в майбутньому для бізнесу, орієнтованого на дані, в якості моделі (інструменту побудови) краще використовувати RNN, а саме мережеву архітектуру LSTM. Архітектура GRU зарекомендувала себе як хороша альтернатива, але через свою інноваційність в деяких випадках вона поступається за якістю роботи. У разі використання архітектури BRNN особливу увагу слід приділити оптимізації процесу навчання мережі (синтезу). Одним із рішень, наприклад, може бути використання нейроеволюційних підходів [30], які добре зарекомендували себе при роботі з RNN. Такі підходи допоможуть значно скоротити час навчання мережі, зберігши при цьому точність прогнозування.

Можливо, в майбутньому, коли буде проведено більше досліджень і вимоги до архітектури GRU стануть більш жорсткими, як це було колись з LSTM, це буде кращим рішенням.

Також важливо звернути увагу на нові підходи до навчання NN, які можуть полегшити та прискорити роботу. Таким чином, процес

технічної діагностики інфокомунікаційних систем значно спроститься.

3.8 Література до розділу 3

1. Miller B. N. Problem Solving with Algorithms and Data Structures Using Python / B. N. Miller, D. L. Ranum, R. Yasinovskyy. – Portland : Franklin Beedle & Associates, 2017. – 438 p.
2. Kryvinska N. Building consistent formal specification for the service enterprise agility foundation / N. Kryvinska // Journal of Service Science Research. – 2012. – Vol. 4, no. 2. – P. 235–269. – DOI: 10.1007/s12927-012-0010-5
3. Kaczor S. It is all about services-fundamentals, drivers, and business models / S. Kaczor, N. Kryvinska // Journal of Service Science Research. – 2013. – Vol. 5, no. 2. – P. 125–154. – DOI: 10.1007/s12927-013-0004-y
4. Kryvinska N. Conceptual Model of Business Services Availability vs. Interoperability on Collaborative IoT-enabled eBusiness Platforms / N. Kryvinska, C. Strauss // Internet of Things and Inter-cooperative Computational Technologies for Collective Intelligence. – Berlin, Heidelberg, 2013. – P. 167–187. DOI: 10.1007/978-3-642-34952-2_7
5. Synergy of Services within SOA / I. Ivanochko [et al.] // Procedia Computer Science. – 2016. – Vol. 98. – P. 182–186. DOI: 10.1016/j.procs.2016.09.029
6. Montavon G. Methods for interpreting and understanding deep neural networks / G. Montavon, W. Samek, K.-R. Müller // Digital Signal Processing. – 2018. – Vol. 73. – P. 1–15. – DOI: 10.1016/j.dsp.2017.10.011
7. ANN-Time Varying GARCH Model for Processes with Fixed and Random Periodicity / E. K. Karuiru [et al.] // Open Journal of Statistics. – 2021. – Vol. 11, no. 05. – P. 673–689. DOI: 10.4236/ojs.2021.115040
8. Fulcher J. Application of Higher-Order Neural Networks to Financial Time-Series Prediction / J. Fulcher, M. Zhang, S. Xu // Artificial Neural Networks in Finance and Manufacturing. – 2006. – P. 80–108. DOI: 10.4018/978-1-59140-670-9.ch005
9. Smith K. A. Neural networks in business: techniques and applications for the operations researcher / K. A. Smith, J. N. D. Gupta //

Computers & Operations Research. – 2000. – Vol. 27, no. 11-12. – P. 1023–1044. DOI: 10.1016/s0305-0548(99)00141-0
10. Tkáč M. Artificial neural networks in business: Two decades of research / M. Tkáč, R. Verner // Applied Soft Computing. – 2016. – Vol. 38. – P. 788–804. DOI: 10.1016/j.asoc.2015.09.040
11. Zhang G.P. Neural networks in business forecasting / G. P. Zhang. – Hershey PA : IRM Press, 2004. – 296 p.
12. Nisbet R. The Data Mining Process / R. Nisbet, J. Elder, G. Miner // Handbook of Statistical Analysis and Data Mining Applications. – Cambridge: Academic Press, 2009. – P. 33–48. DOI: 10.1016/b978-0-12-374765-5.00003-6
13. Tufféry S. Data Mining and Statistics for Decision Making / S. Tufféry. – Chichester, UK : John Wiley & Sons, Ltd, 2011. DOI: 10.1002/9780470979174
14. Multi-step Ahead Time Series Forecasting Based on the Improved Process Neural Networks / Haijian Shao [et al.] // Proceedings of the 9th International Conference on Computer Engineering and Networks. – Singapore, 2020. – P. 397–404. DOI: 10.1007/978-981-15-3753-0_38
15. Vellido A. Characterising and Segmenting the Business-to-Consumer E-Commerce Market Using Neural Networks / A. Vellido, P. J. G. Lisboa, K. Meehan // Progress in Neural Processing. – 2000. – P. 29–54. DOI: 10.1142/9789812813312_0003
16. Limitations of neural networks grow clearer in business [Electronic resource]. – Access mode: https://searchenterpriseai.techtarget.com/feature/Limitations-of-neural-networks-grow-clearer-in-business
17. Neural Networks for Beginners: Popular Types and Applications [Electronic resource]. – Access mode: https://blog.statsbot.co/ neural-networks-for-beginners-d99f2235efca
18. Mitchell D. Using modular neural networks for business decisions / D. Mitchell, R. Pavur // Management Decision. – 2002. – Vol. 40, no. 1. – P. 58–63. DOI: 10.1108/00251740210413361
19. Patan K. Modelling Issue in Fault Diagnosis / Лю Patan // Artificial Neural Networks for the Modelling and Fault Diagnosis of Technical Processes. – Berlin, Heidelberg. – P. 7–27. DOI: 10.1007/978-3-540-79872-9_2

20. Liu, P. Neural Network Evolution Using Expedited Genetic Algorithm for Medical Image Denoising / [P. Liu, Y. Li, M.D. El Basha, R. Fang] // Medical Image Computing and Computer Assisted Intervention (MICCAI 2018) : 21st International Conference, Berlin, Germany, 16-18 September 2018 : proceedings. – Berlin : Nature Springer, 2018. – P.12–20.
21. Capellman, J. Hands-On Machine Learning with ML.NET: Getting Started with Microsoft ML.NET to Implement Popular Machine Learning Algorithms in C# / J. Capellman. – Birmingham : Packt Publishing, 2020. – 296 p.
22. Prospects and Limitations of Non-Invasive Blood Glucose Monitoring Using Near-Infrared Spectroscopy / [J. Yadav, A. Rani, V. Singh, B. Murari] // Biomedical Signal Processing and Control. – 2015. – Vol. 18. – P. 214–227.
23. Babichev, S. Methods, Models and Information Techlology of Complex Data Processing in the Fields of Technical Diagnostics and Bioinformatics / S. Babichev, B. Durnyak. – Lviv : Ukrainian Academy of Printing, 2020. – 180 p.
24. Dow Jones Index Data Set [Electronic resource]. – Access mode: https://archive.ics.uci.edu/ml/datasets/dow+jones+index
25. Online Retail Data Set [Electronic resource]. – Access mode: https://archive.ics.uci.edu/ml/datasets/online+retail
26. Implementation of the Indicator System in Modeling Complex Technical Systems / [S. Leoshchenko, S. Subbotin, A. Oliinyk, O. Narivs'kiy] // Radio Electronics, Computer Science, Control. – 2021. – № 1. – P. 117–126.
27. Neuromodeling of operational processes / [S.A. Subbotin, H.V. Pukhalska, S.D. Leoshchenko, A.O. Oliinyk, Ye.O. Gofman] // Radio Electronics, Computer Science, Control. – 2022. – № 1. – P. 120-129.
28. Neural Network Diagnostics of Aircraft Parts Based on the Results of Operational Processes / [S. Leoshchenko, H. Pukhalska, S. Subbotin, A. Oliinyk, Ye. Gofman] // Radio Electronics, Computer Science, Control. – 2022. – № 2. – P. 69-79.
29. Using Neuromodels for Evaluating and Determining Productivity of Technical Processes / [S. Leoshchenko, O. Nazarenko, A. Oliinyk, S. Subbotin, T. Zaiko, V. Donenko] // International Conference "Problems of Infocommunications. Science and Technology" (PIC

S&T 2020), Kyiv, Ukraine, 6-9 October 2020 : proceedings. – Kharkiv : Kharkiv National University of Radio Electronics, 2020. – P.442-446.
30. Sequencing for Encoding in Neuroevolutionary Synthesis of Neural Network Models for Medical Diagnosis / [S. Leoshchenko, A. Oliinyk, S. Subbotin, T. Zaiko, S. Shylo, V. Lytvyn] // Proceedings of the 3rd International Conference on Informatics & Data-Driven Medicine (IDDM 2020), Växjö, Sweden, 19-21 October 2020. – P. 62-71. – (CEUR Workshop Proceedings, Vol. 2753).

РОЗДІЛ 4
ФРЕЙМВОРКИ ПОБУДОВИ ПРОГРАМНИХ ЗАСОБІВ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ МЕДИЧНОГО ДІАГНОСТУВАННЯ

4.1 Принципи створення фреймворку прийняття рішень для медичного діагностування

На даному етапі існування людства можна виділити достатньо багато різноманітних моделей, які можуть використовуватися для прийняття рішень у відповідних галузях, для розв'язання відповідних проблем. При цьому варто зазначити, що протягом останніх десятиліть відбувся значний зріст кількості таких моделей, що в підсумку призвело до того, що деякі існуючі проблеми були розв'язані вперше, а інші отримали більш якісні розв'язання. Використання засобів машинного навчання на різному рівні призвело до отримання моделей, які ефективно використовуються на практиці.

Однак однією з загальних проблем прийняття рішень на цьому етапі існування людства є те, що такі проблеми в цілому є фрагментованими, відповідно вирішуючи одну з проблем, в кращому випадку відбувається суміжний вплив на іншу проблему. При цьому кожна з таких проблем розглядається окремо, а в підсумку при використанні на практиці взаємний вплив не враховується.

В реальних системах взаємний вплив є важливим і переносити його на певну невизначеність, ризик або оцінювати загально без врахування особливостей невірно. Як відомо, на врахування зв'язків направлена увага в системному аналізі, потреба в якому виникає зокрема тоді, коли проблеми є складними, унікальними. Саме у випадку таких проблем виникає розуміння наявності окремих складових і зв'язків між ними, при чому задіюється міждисциплінарний вплив природи одних складових на інші. Так само і коли розглядаються описані вище проблеми, важливо враховувати їх взаємний вплив на основі методології системного аналізу. Це має сприяти точності отриманих результатів і врахуванню більшої кількості факторів загалом ніж це може бути здійснено, коли будується модель для вирішення однієї відокремленої проблеми.

Відокремлених явищ фактично в природі не існує, тому у випадках, коли розглядаються окремі проблеми, це в підсумку тільки

знижує точність використаної моделі. Це може проявлятися в меншій мірі під час практичного застосування в короткотривалій перспективі, але призводити до значних змін точності у тривалій перспективі, коли традиційний вплив врахованих у моделі факторів змінюється посиленням впливу неврахованих факторів. Таким чином для ефективного розв'язання сучасних проблем прийняття рішень необхідно застосовувати системний підхід для побудови моделей, призначених для розв'язання відповідних проблем.

Однією з галузей людської діяльності, розв'язання проблем в якій формує вплив на інші галузі людської діяльності, є медицина. Вже при створенні діагностичної системи в медицині одразу доводиться зіштовхуватися з тим, що інформація про пацієнта характеризується певним набором даних, який використовується під час різного роду медичних досліджень. Кожне з таких досліджень дозволяє виявляти стан здоров'я пацієнта, при цьому важливим стає і стан пацієнта в минулому.

Звичайно, що не абсолютно всі дані про пацієнта є важливими для будь-якого дослідження, проте нерідко ці фрагменти вибірки повторюються, а в процесі встановлення остаточного діагнозу необхідно провести зазвичай декілька досліджень (оглядів), щоб зрозуміти, який саме діагноз має бути встановлено і яке саме лікування має бути прописано. Тож саме по собі медичне діагностування вже є прикладом системи, в межах якої доводиться розв'язувати декілька підпроблем, але у випадку якщо зв'язок між ними не розглядається, то кожну з них можна розглядати як окрему проблему. Проте на стан пацієнта мають вплив дуже багато різних факторів.

Частина з цих факторів відносяться до тих, що впливають на поточний стан пацієнта. Тобто якщо пацієнт має певну хворобу, то такі фактори призводять до покращення або погіршення його стану.

Інші фактори призводять до того, що у пацієнта з часом формуються певні хвороби. До таких факторів зокрема належать наступні:

– довгостроковий і короткостроковий стан навколишнього середовища (екологічна ситуація в місті загалом і місці перебування людини зокрема), на що здійснює вплив дуже багато факторів, вплив кожного з яких можна розглядати як окрему проблему, що буде розглянуто більш детально нижче;

– індивідуальні фактори, які призводять до впливу на стан здоров'я людини як за короткий, так і за більш тривалий період часу.

До останньої групи факторів, наприклад, можна включити зокрема наступні:

– харчування, куди можуть відноситися як обсяг харчування (тобто фактично недостатнє або надмірне харчування), так і виявлення певних патернів харчування, які з часом можуть призводити до виявлення певних хвороб;

– стресові ситуації, в які потрапляє людина і які призводять до впливу на організм (може розглядатися як окрема задача виявлення стресових ситуацій при моніторингу стану людини через різноманітні трекери фізичного стану або через програмне забезпечення, що дозволяє моніторити робочі процеси працівника);

– загалом умови роботи людини (включаючи зокрема баланс між роботою і відпочинком) тощо.

Звичайно, що цю групу факторів можна значно розширювати. Кожен з таких факторів призводить до впливу, який може виражатися в виявленні деякої хвороби з часом або здійснювати вплив на поточний загальний стан організму, що за послаблення знову ж таки призводить до появи певних хвороб, чи призводить до загострення існуючих хвороб.

У свою чергу до першої групи факторів можна віднести:

– забрудненість повітря: на забрудненість повітря впливає велика кількість процесів, які відбуваються в місці проживання або знаходження людини (наявність виробництва і його вплив, вплив дорожнього трафіку);

– забрудненість води, яку споживає людина, і водоймищ, на яких відпочиває або з якими взаємодіє людина, разом з забрудненістю ґрунту, на якому зокрема вирощуються продукти, які в основному споживає людина;

– вплив сонячного випромінювання тощо.

Загалом варто відзначити, що кількість таких факторів є достатньо великою, вплив частини з них може бути великим, а вплив іншої частини може бути знівельований або оцінений як такий, що не є критичним. Проте як було вказано вище, вплив кожного такого фактору може або прогнозуватися, або знаходитися через виявлення певного патерну поведінки. Прогнозування або виявлення кожного такого впливу є відповідно окремою задачею.

Частина з цих задач розв'язуються окремо, проте загальний вплив враховується в подальшому. Інша частина з цих задач і розв'язується взагалі окремо, і в подальшому враховується тільки як певний фактор, який не оцінений явно.

Розглянемо цю логіку детальніше на більш конкретному прикладі, якому і буде присвячено подальшу частину даного розділу.

Зазвичай задача прогнозування рівня забрудненості повітря розглядається в межах загальних задач, пов'язаних з керуванням містом загалом, або як окрема задача в межах окремого дослідження, що в такому випадку тільки підкреслює відсутність системних зв'язків, які були описані вище. Тоді окремо оцінюється дорожній трафік, його показники в першому випадку використовуються для подальшого планування трафіку в місті. Однак цей же фактор є і фактором, що впливає на здоров'я людей. За довгострокового впливу він може призводити до виявлення у людей хронічних захворювань, а за короткострокового впливу може призводити до впливу на поточний стан здоров'я людей, які вже мають певні захворювання (так само як і довгостроковий). Відповідно комплексний розгляд окремих задач з визначенням впливу їх результатів розв'язання на стан здоров'я пацієнта має призвести до прийняття більш якісних рішень. Такий підхід має також перевести проблему медичного діагностування на більш тривалий строк, який дозволить окрім виявлення вже наявних захворнювань, при цьому маючи більшу кількість даних про пацієнта в історичній перспективі, також створювати певні рекомендації, які дозволять уникнути виявлення хвороб у подальшому.

Задачею даного дослідження не було виявлення всіх наявних факторів, приклади яких були розглянуті вище, а інтеграція цих факторів через розв'язання окремих задач в межах розв'язання загальної задачі. Тому фактично це дослідження сконцентроване на виявленні зв'язків між окремими задачами, факторами, які виявляються в межах задачі прогнозування рівня забрудненості повітря як підзадачі медичного діагностування, яка окрім того, як було описано, є також підзадачею керування містом. Відповідно це говорить про те, що частина факторів з підзадачі прогнозування рівня забрудненості повітря потенційно перетинається з іншими підзадачами, зокрема прогнозування трафіку в місті загалом або на певних ділянках у місті.

Одночасно ця загальна підзадача знаходиться під впливом задачі медичного діагностування, коли загальні оточуючі умови переносяться вже на індивідуальний стан пацієнта. Фактично визначення стану пацієнта є підзадачею, яка розв'язується паралельно описаному вище набору, а в підсумку виноситься рішення про медичне діагностування стану здоров'я пацієнта. При цьому необхідно розуміти, що розв'язати якісно загальну задачу медичного діагностування без розв'язання підзадачі прогнозування рівня забрудненості повітря неможливо. В результаті виходить, що розв'язання цієї підзадачі залежить від розв'язання інших підзадач, які на неї впливають, а вона в свою чергу впливає на розв'язання задачі медичного діагностування.

Фреймворк прийняття рішень для медичного діагностування передбачає поетапне розв'язання підзадач різного рівня з винесенням проміжних рішень та загальних рішень на рівні вже загальної задачі. Важливою його особливістю є врахування поточного стану розв'язання цих задач, а не розгляд загальної задачі відірвано від поточної ситуації. Під поточним станом мається на увазі фактично те, які дані загалом наявні для розв'язання загальної задачі та як вони вже враховуються в межах наявних задач у наявних моделях. Тобто при стандартній постановці наявність зв'язків між факторами може бути врахована і зведена до правильного вибору вхідних ознак. Тоді виявлення всіх множин факторів, а відповідно і підмножин, наявних у певному джерелі, призводить у підсумку до накопичення даних і розв'язання загальної задачі. Але в подальшому це потребує врахування взаємозв'язків між наявними параметрами даних з подальшим визначенням множини ознак, яка буде застосована в моделі. Проте це в підсумку призводить до того, що кожен такий параметр буде врахований декілька разів, а створена в підсумку загальна модель буде характеризуватися певним ступенем надмірності. В залежності від ситуації та обсягу надмірності це може бути критичним фактором. Фактично наявність надмірних моделей призводить до надмірних вимог до ресурсів. Відповідно, якщо результати поточного визначення забрудненості повітря, які наприклад, можуть переглядатися кожну годину, будуть використовуватися при винесенні поточних рекомендацій для пацієнта, то для кожного з наявних пацієнтів кожного разу потрібно буде скористатися відповідними індивідуальними моделями. Тобто

фактично виходить, що в такому випадку весь масив моделей буде задіяний кожен раз, а кількість даних буде постійно розширюватися. Тоді все це буде призводити до необхідності перебудови всіх моделей періодично з накопиченням даних. Відповідно за наявності великої кількості пацієнтів, їх розташування в різних районах міста, наявності великої кількості параметрів, за якими відбувається спостереження, наявності великої кількості станцій для спостереження це може стати практичною проблемою. До того ж варто розуміти, що друга частина моделей, пов'язана з медичним діагностуванням, вже визначає не просто особу за її розташуванням у певному районі, для якого використовуються ті чи інші моделі прогнозування забрудненості повітря, погоди, трафіку, а визначає конкретну особу, а стан її постійно змінюється, відповідно всі побудови, оновлення, використання моделей здійснюються на індивідуальному рівні для кожного пацієнта з певним періодом часу. Цей період часу до того ж має бути щонайменше пов'язаним з тим періодом, який використовується при роботі з першою частиною моделей. Тобто фактично може потребуватися велика кількість ресурсів кожним медичним закладом, що звичайно є практичною проблемою, адже ресурси таких закладів загалом дуже обмежені, у світовому вимірі можна говорити про те, що тільки в окремих випадках здатні задовольнити описані високі вимоги до ресурсів. Тому важливими є при застосуванні фреймворку виявлення наявних моделей, щоб визначити наявні залежності, встановити зокрема, які саме параметри належать до вхідних ознак, які до вихідних, врахування таким чином залежностей, уникнення надмірності при формуванні результуючого набору ознак для кожної моделі, таким чином спрощуючи ті моделі, які ще не створені і які мають бути створені в процесі розв'язання загальної задачі. Це в підсумку дозволяє знизити вимоги до ресурсів, які потрібні для практичного впровадження запропонованих моделей. До того ж частина моделей може бути взагалі реалізована ззовні. У такому випадку ресурси на створення і використання такої моделі взагалі не мають витрачатися.

Відповідно основним результатом використання фреймворку прийняття рішень для медичного діагностування є зменшення вимог до необхідних ресурсів, підвищення ефективності, під якою може розумітися зокрема точність прогнозування моделей за рахунок врахування взаємозалежностей між факторами і зменшення

складності інтеграції наступних моделей у систему або виділення окремих моделей з цієї системи за рахунок можливості відтворення виділених у системі залежностей.

У підсумку загальне представлення фреймворку прийняття рішень для медичного діагностування можна звести до набору наступних етапів.

На першому етапі має бути виділено весь набір ознак $Ft = <ft_1, ft_2, \ldots, ft_N>$, кількість яких дорівнює N і які стосуються результуючої моделі, що має бути створена за результатами аналізу всієї загальної проблеми, тобто це фактично модель медичного діагностування, яка призводить до винесення загального цільового рішення. Це фактично множина потенційних ознак, яка в подальшому має бути зведена до множини Ft'. У загальному випадку $Ft' \subset Ft$, тобто кількість ознак у множині Ft', яка дорівнює N', менше кількості ознак у множині F: $N' < N$. Звичайно, що в певних задачах може виявитися і те, що множини Ft і Ft' ідентичні. У такому випадку це вказує на те, що надмірність моделей відсутня. Тобто це означає, що:

– надмірна залежність між окремими ознаками, яка б призвела до необхідності вилучення однієї з ознак, відсутня;

– усі необхідні значення ознак, що використовуються у загальній вибірці, наявні у вже готових вибірках, тобто немає необхідності прогнозування значення однієї з ознак, значення якої має бути обов'язково представлено або ним не можна нехтувати;

– готові моделі для прогнозування значень ознак з множини Ft відсутні, їх використання є недоцільним або побудова окремих таких моделей (у випадку їхньої відсутності при задоволенні відповідних вимог до таких моделей) вважається недоцільним варіантом. Рішення про недоцільність побудови може бути прийнято, якщо використання вихідних результатів роботи моделі окремо від загальної задачі вважається непотрібним, а ресурсів для реалізації загальної, більш складної, моделі та її функціонування достатньо.

На другому етапі має бути встановлено, які саме моделі $O_q, q = \overline{1, Q}$ та вибірки даних $D_j, j = \overline{1, M}$ охоплюють значення, які можуть використовуватися у складі загальної вибірки даних в якості ознак з множини Ft.

Для виконання даного етапу спочатку має бути встановлено ознаки з множини Ft, значення яких відсутні в потрібному вигляді в наявних

вибірках даних. Відсутність таких даних потенційно може вирішуватися через певні обчислення, які виконуються безпосередньо на основі формул або за допомогою відповідних моделей O_q. Наприклад, такими ознаками можуть бути ознаки, які визначають стан показника на наступний період прогнозування (наступний день, годину тощо). Тоді замість того, щоб використовувати послідовність значень у якості додаткових ознак, щоб фактично спрогнозувати всередині моделі значення на майбутній період та визначити його вплив на значення іншої ознаки, яка є основним виходом (вихідною ознакою) побудованої моделі, в той самий майбутній період, використовується одне значення, яке визначає конкретний вплив даної ознаки в майбутній період. Таким чином формується множина моделей *O*, яка в подальшому використовується для створення загальної множини моделей для розв'язання загальної задачі. Ці моделі або вже існують на даний момент, при чому вони можуть бути у розпорядженні сторони, яка розв'язує загальну задачу, або вони можуть належати іншій стороні, надаючи тільки результат застосування моделі, або мають бути створені для зменшення необхідних загальних ресурсів через уникнення надмірності, як це було описано вище.

Далі має бути виконано дослідження взаємозалежності між ознаками з множини *Ft*. Це фактично вибір ознак для моделей, які застосовуються у складі фреймворку. Якщо такі взаємозалежності існують і фактично значення певних ознак є надмірними, то в підсумку має бути виділено підмножину ознак Ft'. У процесі цього виділення може бути скорочено і множину моделей *O*, тобто фактично виділено в підсумку тоді певну множину моделей O'.

Фактично виділення множини моделей O', підмножини ознак Ft' та відповідних цим ознакам вибірок даних, що формують загальну вибірку (множину вибірок) *D,* є результатом виконання даного етапу.

На третьому етапі має бути реалізовано роботу над обраними моделями, які потребують власної участі сторони, яка відповідальна за прийняття рішень в межах фреймворку прийняття рішень для медичного діагностування. Тобто моделі з множини O', які не потребують створення, а вже є створеними і навченими, не розглядаються на цьому етапі.

Під час реалізації етапу спочатку відбувається на основі характеристик кожної з моделей O' формування навчальних множин з загальної вибірки (множини вибірок) *D*. Далі виконується безпосередньо створення тих моделей, які ще не створені, а після цього виконується навчання цих моделей за допомогою підготовлених навчальних вибірок.

На четвертому етапі відбувається застосування створених моделей і моделей, які раніше існували, з множини O'. У підсумку множина моделей призводить до поточного розв'язання задачі прийняття рішень для медичного діагностування. Дана задача розв'язується динамічно, тобто зі встановленим періодом потребує постійного прийняття рішень. У процесі реалізації даного етапу періодично може виконуватися оновлення моделей з множини O' на основі нових отриманих даних.

Відповідно, застосовуючи даний фреймворк, потрібно далі визначити потрібні моделі, ознаки згідно з тими підзадачами, які були описані вище. При цьому в межах даної роботи фреймворк прийняття рішень для медичного діагностування буде формуватися і застосовуватися в частині отримання загальних даних (не індивідуальних), тобто фактично отримання даних про рівень забрудненості повітря. У подальшому це може бути розширено, але вже ця задача дозволяє розглянути і використати принципи фреймворку, які в загальному вигляді були описані в даному підрозділі. При цьому враховуючи необхідність визначення загальних принципів, які можуть потім використовуватися на практиці для різних підприємств, підзадач, міст, моделі з множини O' будуть створені в роботі, а не використані певні готові моделі, але на практиці в залежності від наявності або відсутності таких моделей можуть бути прийняті відповідні рішення.

Відповідно логіка застосування фреймворку, описаного в даному підрозділі, в межах експериментального дослідження полягає фактично у прогнозуванні автомобільного трафіку на певних станціях міста, що в подальшому використовується для прогнозування рівня забрудненості атмосферного повітря як значення вхідних ознак, а отримані результати в свою чергу стають основою порад пацієнту щодо його дій в умовах відповідного рівня забрудненості повітря в наступні години.

4.2 Дослідження математичних моделей для прогнозування автомобільного трафіку

Прогнозування автомобільного трафіку є важливою практичною задачею, яка загалом полягає у визначенні трафіку, тобто кількості транспортних засобів, які долають певну ділянку місцевості протягом заданого періоду часу в майбутньому на основі даних про рівень трафіку в минулому, а також на основі даних про інші показники, якщо такі дані є релевантними для визначення відповідної моделі.

У даному підрозділі розгляд задачі прогнозування автомобільного трафіку відбувається в межах години, тобто кожна характеристика екземпляру вибірки даних відповідає значенню, яке було накопичено протягом години. Це передбачає рух транспортних засобів у всі боки: тобто якщо контроль здійснюється на ділянці руху транспорту, яку транспорт долає в двох напрямках (фактично назустріч реєстратору та в протилежному напрямку), то рівень трафіку обчислюється як сума кількості транспортних засобів, які подолали цю ділянку в одному напрямку, та кількості транспортних засобів, які подолали цю ділянку протилежному напрямку. Якщо фіксація здійснюється на перехресті, то відповідно обчислюється сума кількості транспортних засобів зі всіх напрямків. Це дозволяє визначити активність транспортних потоків. При цьому якщо певний транспортний засіб подолав деяку ділянку за годину в декількох напрямках або в одному повторно, то він враховується декілька разів.

Задачу прогнозування автомобільного трафіку тоді можна розглядати як визначення трафіку на деякій станції A протягом кожної з H^F годин у майбутньому на основі даних про рівень трафіку протягом попередніх H^P годин за цією станцією A, а також за іншими станціями з множини B та на основі даних за іншими показниками з множини E:

$$tr_A^{t+h} = f\begin{pmatrix} tr_A^t, tr_A^{t-1}, \dots, tr_A^{t-H^P}, tr_b^t, tr_b^{t-1}, \dots, tr_b^{t-H^P}, \dots, \\ v_e^t, v_e^{t-1}, \dots, v_e^{t-H^P} \end{pmatrix}, \quad (4.1)$$
$$h = \overline{1, H^F}, b \in B^S, B^S \subseteq B, e \in E,$$

де tr_A^{t+h} – величина трафіку на станції A протягом $(t+h)$-ої години (з $t+h$ годин 0 хвилин до $t+h$ годин 59 хвилин) у

майбутньому;

$tr_A^t, tr_A^{t-1}, \dots, tr_A^{t-H^P}$ – рівень трафіку за станцією A протягом t -ої, $(t-1)$-ої, $(t-H^P)$-ої годин у минулому;

f є функціональною залежністю, яку необхідно знайти в результаті створення моделі прогнозування автомобільного трафіку для станції A;

$tr_b^t, tr_b^{t-1}, \dots, tr_b^{t-H^P}$ – рівень трафіку за станцією b протягом t -ої, $(t-1)$-ої, $(t-H^P)$-ої годин у минулому;

B – множина станцій, за якими здійснюється вимірювання трафіку в місті або станцій, вимірювання за якими доступні для побудови моделі;

B^S – підмножина станцій, яку було обрано з множини B на основі визначення ступеня впливу трафіку за ними на трафік за станцією A;

E – множина показників, які здійснюють вплив на значення рівня трафіку за станцією A протягом кожної з $(t+h)$-х годин;

$v_e^t, v_e^{t-1}, \dots, v_e^{t-H^P}$ – значення показника e з множини E протягом t -ої, $(t-1)$-ої, $(t-H^P)$-ої годин у минулому;

H^F – кількість наступних годин від поточного моменту часу, для якої виконується прогнозування;

t – номер поточної години, тобто момент часу, в який виконується прогнозування;

H^P – кількість попередніх годин від поточного моменту часу *t*, на основі якої виконується прогнозування, не враховуючи поточну годину, для якої трафік вже відомий (або інакше кількість годин, на основі якої відбувається прогнозування, дорівнює H^P+1).

Множина показників E фактично складається з підмножини відомих показників E^K, за якими такий вплив встановлено і підтверджено, та підмножини невідомих показників E^U, які фактично при створенні моделі визначаються як вплив певної невизначеності. При створенні моделі для розв'язання задачі (4.1) необхідно зменшувати підмножину E^U за рахунок збільшення підмножини E^K, однак врахувати всі фактори, які впливають на вихідну ознаку за сучасного рівня науки неможливо, тому вплив величин показників з підмножини E^U існує і призводить до того, що вихідне значення, розраховане на основі створеної моделі не співпадає повністю зі значенням, отриманим на основі спостережень у майбутньому.

У даному підрозділі задача (4.1) розглядається в аспекті прийняття рішень для медичного діагностування, тому визначення періоду часу, який формується встановленням значення H^F, виконується на основі врахування відповідного періоду в задачі прогнозування рівня забрудненості атмосферного повітря. Тому в даному підрозділі прогнозування відбувається для 6 наступних годин. При цьому значення величини H^P має бути визначено на основі проведених експериментів. Необхідно зокрема перевірити, як впливає збільшення періоду використовуваних історичних даних на точність прогнозування. Зокрема це стосується того, чи призводить до збільшення точності прогнозування збільшення величини H^P до значення, більшого ніж H^F. Окрім того в процесі експериментальних досліджень необхідно також встановити, яким чином необхідно обрати підмножину B^S з множини B.

Означені завдання є завданнями експериментального дослідження, але перед тим як його проводити, було проведено аналіз існуючих методів і моделей, які на даний момент використовуються для розв'язання задачі (4.1) в певних її варіантах постановки. При цьому зважаючи на подальше застосування створеного рішення, розглядались саме ті роботи, які стосуються вирішення проблеми в короткостроковому періоді, адже ця задача може розглядатися в певних модифікаціях і в довгостроковому періоді. Тоді прогнозування здійснюється, наприклад, для розв'язання задачі планування в місті, коли необхідно визначити транспортні потоки для подальшого стратегічного планування діяльності в місті.

Проведений аналіз наявних результатів виконувався на основі цілого ряда джерел з врахуванням результатів, представлених у огляді літератури за загальною проблемою [1].

Отримані результати дозволяють стверджувати, що велика частина робіт на даний момент направлена на об'єднання моделей на основі глибоких нейронних мереж та моделей, які дозволяють враховувати співвідношення між вхідними даними, зокрема взаємне розташування відповідних станцій збору даних (датчиків). Такий підхід дозволяє в підсумку створити модель, яка дозволяє опрацьовувати дані зі всіх станцій одночасно і робити прогнози щодо їх подальших станів. Він розглядається зокрема в наступних дослідженнях.

У роботі [2] запропоновано використовувати комбіновану модель на основі графової згорткової нейронної мережі, яка витягає характеристики топологічної структури з даних трафіку, Long Short-Term Memory (LSTM), яка витягає характеристики часової структури, та залишкової нейронної мережі, яка використовується для оптимізації загальної моделі. Безпосередньо експериментальне дослідження використовує на вхід моделі дані за 5, 15 або 30 хвилин, при цьому для оцінки результатів використовуються показники середньої абсолютної похибки або Mean Absolute Error (MAE), середньоквадратичної похибки або Mean Squared Error (MSE) та коефіцієнт R^2, отримані значення яких знижуються зі збільшенням відповідного періоду. Експериментальні дослідження виконані на даних з Каліфорнії (США) з системи, яка передбачає 39000 датчиків.

У роботі [3] створено гібридну графічну модель , яка відрізняється від зокрема попередньої тим, що окрім статичного графіку в основі передбачає також створення динамічного графіку, що дозволяє не тільки безпосередньо представити топологію всієї мережі трафіку, але і оновлювати наявну інформацію, таким чином відповідаючи реальним умовам, в яких реалізується трафік. Проте сама модель у даній роботі поєднує використання графової нейронної мережі, згорткової нейронної мережі та передбачає використання механізму уваги для виділення в підсумку просторово-часових характеристик. Експериментальне дослідження відбувається на 2 вибірках даних, розташованих у публічному доступі, що ґрунтуються на даних, зібраних у місті Лос Анджелес (США) з 1500 датчиків та у Каліфорнії (США). Оцінювання результатів відбувалось на основі показників MAE, кореня середньоквадратичної помилки або Root Mean Square Error (RMSE) та середньої відносної помилки або Mean Absolute Percentage Error (MAPE).

Автори у роботі [4] пропонують використання просторово-часової графової згорткової нейронної мережі. Дослідження відбувалось на основі вхідних даних за 60 попередніх хвилин з прогнозуванням трафіку за наступні 15, 30 або 45 хвилин. При цьому отримані результати як для даних, зібраних у Каліфорнії, так і для даних, зібраних у Пекіні, значно відрізняються за показниками MAE, MAPE, RMSE зі збільшенням періоду прогнозування: фактично результати погіршуються більше ніж у 1,5 рази при збільшенні періоду прогнозування з 15 до 45 хвилин.

Окрім того існує ряд досліджень, пов'язаних з прогнозуванням трафіку, але в ракурсі інших показників. Зокрема у роботі [5] відбувається прогнозування швидкості руху, що звичайно не є задачею даного підрозділу, але при цьому також використовуються подібні структури до розглянутих. Модель створюється на основі згорткової довгострокової пам'яті (Conv-LSTM), механізму уваги та двох двонаправлених LSTM (BiLSTM). Таке рішення підкреслює розповсюдженість описаних моделей для фактично моделювання трафіку та прогнозування в подальшому, але вже на основі іншого показника.

Однак, в розрізі загальної проблеми прийняття рішень для медичного діагностування такий підхід потребує з одного боку наявності відповідних обчислювальних потужностей у медичному центрі, адже моделі, які створюються, передбачають достатньо складну структуру, а з іншого боку – доступу до таких даних, що не всюди є в наявності. При цьому відсутність доступу може бути обумовлена не тільки відсутністю доступу саме певного медичного центру, а відсутністю таких даних за містом загалом. Тобто в країнах, які розвиваються, дані щодо трафіку в містах є достатньо фрагментованими, не охоплюють всі можливі вулиці та перехрестя або певну їх підсистему, відповідно створити цілісну картину за допомогою таких моделей достатньо проблемно. Тоді логіка просторових співвідношень порушується через відсутність частини позицій у загальній мережі доріг. Окрім того дані можуть бути не доступні протягом певного, часом достатньо тривалого періоду. Наприклад, у зв'язку з військовим станом такі дані не надаються і відсутні через загальнодоступні сервіси. До того ж у розрізі проблеми прийняття рішень для медичного діагностування вцілому та прогнозування рівня забрудненості атмосферного повітря зокрема потреба прогнозування трафіку на кожній ділянці в місті може бути відсутня. Зокрема через те, що і станції для збору даних щодо рівня забрудненості атмосферного повітря розташовані тільки на окремих позиціях, їх кількість є обмеженою. Тож у підсумку можна розглядати задачу (4.1) в умовах, коли множина станцій B є значно обмеженою. І тоді підхід, представлений в роботах [2]-[4] у даному розрізі стає відповідно неактуальним. При цьому слід також відзначити, що результати приведених досліджень та інших проаналізованих робіт вказують на те, що прогнозування автомобільного трафіку в більш

тривалій перспективі (більше години) є окремою задачею і потребує докладного дослідження для покращення результатів.

Окрім того проведені в розглянутих роботах дослідження ґрунтуються на однакових вибірках, що в підсумку призвело до того, що в роботі [1] відповідні результати за різними роботами були зведені в єдину таблицю. Таке порівняння звичайно є важливим, однак ці вибірки даних характеризуються великою кількістю датчиків, що не відповідає описаним умовам, в яких в даній роботі розв'язуються відповідні задачі.

Зробивши дані висновки, розглянемо також ряд робіт, які не враховують просторові співвідношення між станціями спостереження, а тому можуть бути використані потенційно для розв'язання задачі (4.1) в умовах обмежених даних про трафік у місті. Даний напрямок досліджень було виділено на основі робіт [6]-[9], однак загалом він є значно ширшим.

У роботі [6] проведено експериментальне дослідження, де фактично здійснюється прогнозування трафіку на наступні 5 хвилин за даними попередніх 30 хвилин. Дослідження проводилось на основі використання моделі AutoRegressive Integrated Moving Average (ARIMA) та моделей штучних нейронних мереж з архітектурою LSTM та Gated Recurrent Unit (GRU). Для дослідження використано дані, зібрані з 15000 датчиків у Каліфорнії, однак для тестування випадковим чином вибиралось лише 50 датчиків. Результати дослідження оцінювались на основі показників MSE та MAE. Отримані результати продемонстрували перевагу LSTM та GRU моделей порівняно з ARIMA.

Автори роботи [7] виконали дослідження математичних моделей, які можуть застосовуватися для розв'язання проблеми прогнозування автомобільного трафіку на основі даних, зібраних у місті Пекін (Китай) з 500 станцій. Дослідження охоплювало серед використаних моделей ARIMA, метод опорних векторів, радіально-базисну нейронну мережу, стековий автокодувальник, ванильну рекурентну нейронну мережу та LSTM. Прогнозування виконувалось на 15, 30, 45, 60 хвилин. Отримані результати для різних моделей погіршувались з часом, проте для LSTM залишались досить стабільними. Окрім того моделі на основі LSTM продемонстрували надійні результати при проведенні експериментів, що дозволило їх, як і в багатьох інших дослідженнях рекомендувати як найкращий

базовий варіант використання. Також використання LSTM-моделей апробовано зокрема в роботі [8] та цілому ряді інших робіт.

У роботі [9] приділяється увага використанню саме двонаправлених LSTM (BiLSTM) для короткострокового прогнозування трафіку на основі ряду показників, включаючи завантаженість, швидкість. Однак, дане дослідження проведено на основі даних, отриманих шляхом моделювання для автостради у Мельбурні (Австралія). Таке дослідження є важливим, однак не дозволяє отримати результати на реальних даних. Відповідно шляхи подальшого використання швидше полягають в моделюванні загальних транспортних потоків у певних умовах для прийняття рішень щодо керування цими потоками, а не щодо прийняття рішень на основі точково прогнозованих значень, які визначають процеси, що реально відбуваються на певній місцевості. Прогнозування виконувалось на період до 60 хвилин (5, 10, 15, 30, 45, 60): зі збільшенням інтервалу прогнозування зменшувалась точність створених моделей. В якості вхідних даних використовувались дані одного з показників за минулий період за станцією, для якої відбувається прогнозування.

Роблячи висновки за другою групою робіт, варто вказати на те, що загалом вони демонструють придатність LSTM-моделей для розв'язання задачі прогнозування автомобільного трафіку на високому поточному рівні. Однак дослідження найчастіше роблять акцент лише на застосуванні таких моделей на основі даних про значення прогнозованого показника за минулий період на тій самій станції. Це значно відрізняється від логіки першої групи методів, де фактично використовуються дані групи станцій для врахування просторових співвідношень. При цьому помітним є те, що перша група робіт враховує дані всіх станцій, а друга не використовує дані в наявному обсязі. Тобто відбувається створення окремих моделей для окремих станцій, але не досліджується, чи не є релевантними для створення цих окремих моделей дані деяких інших станцій. Тоді перший підхід використовує всі такі станції, а другий – жодні.

У даному розділі для визначеної задачі (4.1) виконано експериментальне дослідження, для якого використано вибірку даних, побудовану на основі даних, які були зібрані в місті Мадрид (Іспанія) та які розміщені у вільному доступі на Порталі відкритих даних Мадридської міської ради [10], включаючи наступні окремі вибірки:

– вибірка даних стосовно автомобільного трафіку в місті Мадрид, що включає погодинні дані з 1 січня 2018 року до 30 вересня 2022 року [11];

– вибірка даних стосовно погодних умов у місті Мадрид, що включає погодинні дані з 1 січня 2019 року до 30 вересня 2022 року [12].

Саме погодні умови розглядались під час експериментального дослідження стосовно зменшення підмножини E^U, щоби збільшити точність побудованої моделі в підсумку. Використання цих показників призвело до того, що простір використовуваних даних був зведений до підвибірки, яка охоплює дані з 1 січня 2019 року до 30 вересня 2022 року. Розглядаючи дане звуження, варто також акцентувати увагу на тому, що горизонт наявних даних впливає на складність тих моделей, які можуть використовуватися. Якщо прогнозування відбувається не у короткому періоді (зазвичай від 5 хвилин), а на години, то це призводить до того, що в результаті створення часового ряду кількість екземплярів вибірки зменшується. При цьому тим активніше, чим більшим є період прогнозування. При цьому підвищення складності моделі, а збір значень різних показників за певну кількість годин за всіма станціями також слід вважати таким ускладненням, варто враховувати також і таким чином, що це впливає на обсяг необхідних даних. Тоді навіть у випадку, якщо кількість станцій для збору даних про автомобільний трафік збільшується в певний момент часу, це все одно потребує в подальшому накопичення таких даних протягом значного інтервалу часу, що не дозволяє використовувати більш складні моделі в таких випадках.

Спочатку було проаналізовано дані вибірки, що характеризує трафік у місті Мадрид. Вона включає дані з 1 січня 2018 року до 30 вересня 2022 року, зібрані за 60 станціями в місті. За кожною станцією спостереження відбувалось у двох напрямках руху. У розрізі даної задачі, враховуючи подальше застосування для визначення забрудненості повітря, будуть використовуватися об’єднані дані, тобто за кожною станцією зафіксовано потік у два боки сумарно за кожну годину. За результатами аналізу даних за період з 1 січня 2019 року до 30 вересня 2022 року було виявлено, що дані за станцією Calle Arenal відсутні, тому її було виключено з подальшого дослідження, яке проводилось за 59 станціями.

Кожен окремий файл представляє зібрані за трафіком дані за місяць. У кожному файлі кожен окремий рядок складається з:

– запису, що відповідає даті, за якою було зібрано дані;

– номеру станції;

– додаткової позначки (перша половина дня з прямим напрямком руху, перша половина дня зі зворотнім напрямком руху, другу половина дня з прямим напрямком руху та другу половина дня зі зворотнім напрямком руху);

– послідовність з 12 стовпців, кожен з яких містить трафік за відповідну годину за відповідною станцією на відповідну дату.

Дані було програмно агреговано і збережено у сформованій вибірці даних Вибірка даних містить окремо дані за кожною станцією. Відповідно для кожної станції кожен екземпляр даних визначає:

– дату і час, в які було зібрано дані;

– кількість транспортних засобів, які подолали дану станцію сумарно в прямому та зворотному напрямках: для годин до 12 це сума кількості транспортних засобів за першу половину дня для зворотного та прямого напрямків за відповідну годину, а для годин більше 12 – відповідно сума кількості транспортних засобів за другу половину дня для зворотного та прямого напрямків за відповідну годину.

Так само в файлах помісячно зберігаються дані про погоду в місті. Кожен такий файл містить рядки, що складаються з:

– провінції;

– муніципалітету;

– номеру станції;

– показника, значення якого вимірюється;

– узагальненого представлення станції у вигляді коду, що включає провінцію, муніципалітет тощо;

– року спостереження;

– місяця спостереження;

– дня спостереження;

– 48 стовпчиків, які послідовно містять 24 пари, що відповідають годинам від 1 до 24: кожна пара складається з 2 стовпчиків, перший з яких містить значення відповідного показника за відповідну годину, а другий – позначку про коректність внесених даних.

Якщо внесені дані некоректні, то до сформованої вибірки вносилося значення NaN, яке в подальшому необхідним чином оброблялось, що буде представлено нижче.

Вибірка містить значення для наступних показників:
– швидкість вітру;
– напрямок вітру;
– температура атмосферного повітря;
– відносна вологість;
– атмосферний тиск;
– сонячне випромінювання;
– кількість опадів.

На основі цих даних було проведено експериментальне дослідження прогнозування автомобільного трафіку за допомогою використання нейронних мереж на основі архітектури LSTM. Дана архітектура відзначається здатністю враховувати залежності майбутніх значень прогнозованих показників від значень у минулому. Саме такою за своєю структурою і є задача (4.1). Однак вибір архітектури LSTM у подальшому визначає необхідність налаштування параметрів моделі.

Тоді для прогнозування автомобільного трафіку необхідно для кожної станції побудувати модель, яка на основі вхідних ознак буде прогнозувати значення вихідних. У даному випадку, зважаючи на особливості загальної задачі прийняття рішень для медичного діагностування та задачі прогнозування рівня забрудненості атмосферного повітря, прогнозування має здійснюватися не на 1 годину вперед, а на 6 годин. Прогнозу на 1 годину буде достатньо тільки для короткострокове попередження пацієнтів про загрозу їхньому здоров'ю, на яку фактично не всі можуть мати змогу зреагувати. У кращому випадку тоді можна використати засоби індивідуального захисту, але змінити маршрути пересування, плани в багатьох випадках буде важко, тому вплив сформованих рішень буде фактично низьким у багатьох випадках. Тому в даній роботі для побудови системи медичного діагностування було вирішено виконувати прогнозування на період у 6 годин у майбутньому. Це має дозволити пацієнтам врахувати в середньостроковій перспективі зміну умов, при цьому залишаючи такі прогнози достатньо точними, чого і необхідно прагнути в даній роботі у підсумку. Звичайно, що пацієнт і лікар не отримують дані про рівень трафіку, а вони використовуються у складі описаного в підрозділі 4.1 фреймворку. Тому дані про трафік протягом кожної з 6 наступних годин повинні

використовуватися на вхід моделі для прогнозування рівня забрудненості атмосферного повітря за 6 наступних годин.

При цьому для подальшого дослідження залишається окрім вибору внутрішніх параметрів LSTM-моделі відкритим питання вибору величини H^P та способу, яким необхідно обрати підмножину станцій B^S з множини станцій B.

Тому для подальших експериментів було створено декілька вибірок даних на основі часових рядів, що в подальшому були задіяні для порівняння. У результатах у подальшому винесено результати для вибірок, які базуються на 6 попередніх годинах та на 24 (прогнозування наступних 6 годин за минулою добою). Кожна така вибірка була організована наступним чином.

Результуюча структуру, яка і використовувалась у подальшому для проведення самих експериментів, представлена словником, у якому за ключами, що відповідають коду станції спостереження, було внесено дані стосовно кожної відповідної станції. Дані за кожною станцією представлені словником, у якому за ключами доступні:

– дані вхідних ознак для проведення навчання;

– дані вихідних ознак, необхідні для проведення навчання;

– дані вхідних ознак для проведення тестування;

– дані вихідних ознак, необхідні для проведення тестування;

– дати (дата і час), які відповідають кожному екземпляру, що використовується під час навчання;

– дати (дата і час), які відповідають кожному екземпляру, що використовується під час тестування.

Для формування цих даних відносно кожної станції було виконано наступну процедуру.

Спочатку було сформовано структуру, рядки якої відповідають послідовно кожній годині в проміжку часу від 1 січня 2019 року до 30 вересня 2022 року. Далі за кожним рядком було внесено відповідне значення з раніше сформованої на основі файлової структури вибірки даних стосовно відповідної станції, для якої дані формувались. Якщо значення було відсутнє, то вносилось значення NaN. Далі була виконана передобробка даних. Для цього було визначено всі позиції, для яких встановлено значення NaN, і для кожної такої позиції було виконано пропорційне визначення пропущеного значення, тобто встановлено попереднє та наступне наявні значення, а далі рівномірно

розподілено зміну між цими значеннями на кількість пропущених позицій.

На наступному кроці було виконано нормалізацію отриманих значень. Для цього було спочатку визначено кількість екземплярів, які мають бути віднесені до навчальної вибірки. Навчальна вибірка формувалася за принципом виділення 80 % екземплярів початкової вибірки даних. Інші 20 % відповідно відносилися до тестової вибірки. При цьому було враховано розмір партії екземплярів для навчання. Розмір даної партії (параметр BATCH_SIZE) було встановлено на рівні 32. Для коректності наступних процедур відповідно було узгоджено кількість екземплярів у навчальній вибірці з даним параметром: якщо кількість не були кратною 32, то кількість цих об'єктів зменшувалась шляхом приєднання до тестової вибірки таким чином, щоб результуючий розмір став кратним 32. Далі за виділеною таким чином підмножиною екземплярів було визначено мінімальне і максимальне значення на тренувальній вибірці, після чого безпосередньо виконано нормалізацію даних для приведення вхідних даних до інтервалу від 0 до 1. Нормалізація виконувалась тільки на основі параметрів навчальних даних для того, щоб не спотворювати роботи створеної в результаті мережі. Тобто у випадку отримання більших або менших значень у подальшому (під час тестування) модель не має інформації про такі значення спочатку, адже вони відсутні в навчальній вибірці.

Після цього було створено структуру, кількість стовпчиків якої дорівнює $H^F + H^P$. Кожен стовпчик заповнено сформованими на попередньому кроці значеннями шляхом копіювання. Далі було виконано зсув цих даних зі збільшенням зсуву в кожному наступному стовпчику на -1. Цю структуру було занесено в словник, у якому за ключами зберігаються безпосередньо ці значення та окремо значення дати зі зсувом на $-H^P$. Таким чином було отримано дату і час, для яких має виконуватися прогнозування за кожним екземпляром вибірки.

Далі ці дані були розподілені на:

– дані вхідних ознак для проведення навчання через виділення H^P перших значень (стовпців) приблизно 80 % перших екземплярів, як було описано вище;

– дані вихідних ознак, необхідні для проведення навчання через виділення H^F останніх значень (стовпців) приблизно 80 % перших екземплярів, як було описано вище;

– дані вхідних ознак для проведення тестування через виділення H^P перших значень (стовпців) приблизно 20 % останніх екземплярів, як було описано вище;

– дані вихідних ознак, необхідні для проведення тестування через виділення H^F останніх значень (стовпців) приблизно 20 % останніх екземплярів, як було описано вище.

Таким чином сформовані дані було занесено у загальний словник за ключем відповідної станції для кожного відповідного ключа з додаванням також за окремими ключами переліку дат, які відповідають екземплярам з навчальної вибірки та окремо тестової вибірки. Також за окремими ключами було внесено дані про мінімальне та максимальне значення за навчальною вибіркою.

При цьому за вхідними ознаками (перший та третій пункт представленого вище переліку) було сформовано тривимірний масив. Як було описано вище, сформований масив є двомірним. Відповідно від розташовувався у тривимірному масиві за індексом 0. За всіма іншими індексами було додано метеорологічні дані. Тобто кожен індекс – це один з перелічених вище параметрів, що описують погоду. Оскільки існує ціла множина станцій, які вимірюють дані показники, то процедура вибору виконувалась наступним чином за кожним показником окремо.

Спочатку, перед виконанням дій за кожним показником окремо, було обчислено відстань між кожною станцією, яка виконувала метеорологічні вимірювання (погоди), та кожною станцією, яка виконувала вимірювання трафіку. За кожною станцією, яка виконувала вимірювання трафіку, було сформовано перелік станцій, які виконували метеорологічні вимірювання, відсортувавши їх за відстанню від станції, для якої цей перелік створювався.

Далі для кожного показника за відповідної станції, яка виконувала вимірювання трафіку, у порядку збільшення відстані розглядалися послідовно станції, які виконували метеорологічні вимірювання. Якщо дана станція вимірювання за цим показником виконувала, то її результати заносились до результуючого набору екземплярів. Якщо ні, то відбувався перехід до наступної станції. Як тільки дані були знайдені, відбувалося формування часового ряду за описаними вище

принципами, так само спочатку заповнюючи пропущені значення, потім нормуючи дані, повторюючи на необхідну кількість стовпців, але вже в кількості H^P, а тоді зсуваючи кожен з них на номер індекса.

Після того, як було заповнено всі вхідні ознаки, процес завершувався, а сформована вибірка використовувалась далі під час проведення досліджень.

У процесі дослідження було проведено ряд експериментів стосовно наступних моделей:

– LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку;

– biLSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку;

– LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку та метеорологічними даними, співвіднесеними з цією станцією за цей же період часу;

– LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку, та станціями, які визначені як релевантні для станції прогнозування;

– LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 24 години за станцією, за якою відбувається прогнозування трафіку;

– biLSTM-моделі з вхідними ознаками на основі даних трафіку за минулу оптимальну (вибрану відповідним чином) кількість годин за станцією, за якою відбувається прогнозування трафіку, та станціями, які визначені як релевантні для станції прогнозування.

Далі описано параметри використаних моделей. Для вибору конкретних значень проводились додаткові експерименти. Найкращі результати були визначені для подальших рішень. Додаткові варіанти моделей, які не призвели до отримання кращих результатів, не приводились у підсумкових результатах. Окрім того були вилучені і LSTM-моделі з використанням метеорологічних даних в якості вхідних ознак, оскільки вони не призвели до покращення відносно моделей, які ці ознаки не враховували. Тобто в підсумку такі моделі

тільки мали складнішу структуру, що загалом не відповідає умовам розв'язання задачі (4.1), які були описані раніше.

Для оцінювання отриманих результатів прогнозування було визначено набір критеріїв, що включають:

– MSE;
– MAE;
– RMSE;
– R^2.

Базові LSTM-моделі створювались на основі внутрішньої структури з 1, 2 та 3 шарами LSTM-чарунок.

У підсумку було використано наступну структуру LSTM-моделі:

– вхідний шар, який складається з 6 ознак (значення трафіку за минулі 6 годин);
– 32 LSTM-чарунки на першому внутрішньому шарі;
– виключення (dropout) у 0,1;
– 32 LSTM-чарунки на другому внутрішньому шарі;
– виключення (dropout) у 0,1;
– повнозв'язний шар з 6 нейронами для формування остаточних результатів.

Для оптимізації моделі було використано оптимізатор Адама. В якості функції витрат для регуляції навчання використано MSE. Навчання максимально реалізовувалось протягом 500 ітерацій, але при цьому було встановлено критерій раннього зупину для уникнення перенавчання моделей. Якщо протягом максимум 40 ітерацій не відбувалося покращення результатів навчання, що полягало в зменшенні значення функції витрат, то навчання завершувалось.

Приклад графіків, що відображають процес навчання моделі описаної архітектури та структури, представлено на рис. 4.1 для станції Calle Sinesio Delgado та на рис. 4.2 для станції Calle Hermanos Garcia Noblejas. На цих графіках синьою лінією показано результати на основі навчальної вибірки, а помаранчевою – валідаційної. Для цього сформована, як було описано вище, навчальна вибірка була розділена на частину, яка використовувалась для безпосередньо навчання, – 75 % від початкової навчальної вибірки (60 % від загальної вибірки даних), та 25 % – на валідаційну (20 % від загальної вибірки даних).

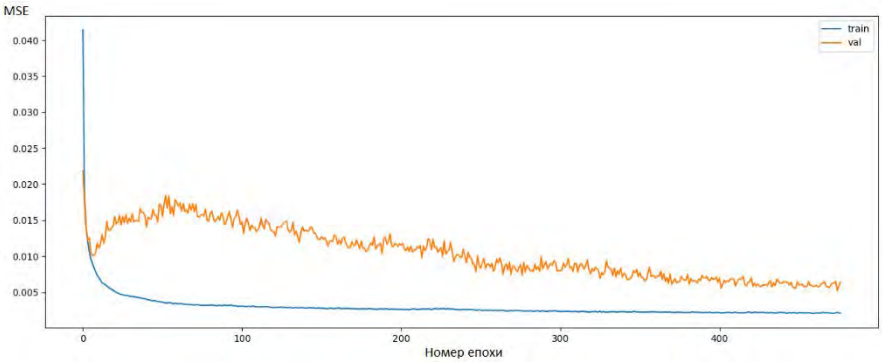


Рисунок 4.1 – Зміна значення функції витрат на основі показника MSE протягом навчання LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією Calle Sinesio Delgado

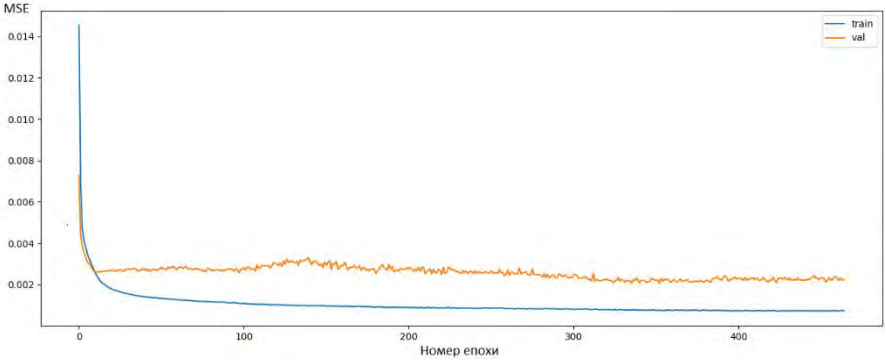


Рисунок 4.2 – Зміна значення функції витрат на основі показника MSE протягом навчання LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією Calle Hermanos Garcia Noblejas

У таблиці 4.1 приведено повні результати навчання та тестування створених моделей на основі даної архітектури та структури. У стовпчику навчання представлено результуюче значення MSE, яке було отримано в процесі навчання за моделлю на навчальній вибірці.

Таблиця 4.1 – Результати навчання та тестування LSTM-моделей з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку

Станція	Навчання	MSE	MAE	RMSE	R^2
1	2	3	4	5	6
Paseo de la Castellana	0,00082	0,00208	0,02971	0,04566	0,85122
Calle Princesa	0,00201	0,00591	0,0531	0,07689	0,74462
Calle Doctor Esquerdo	0,00225	0,00715	0,04936	0,08455	0,86076
Paseo de San Franciso de Sales	0,00033	0,00069	0,01692	0,02629	0,82379
Paseo de Santa Maria de la Cabeza	0,00085	0,00178	0,02954	0,04214	0,88285
Calle Arturo Soria	0,00019	0,00052	0,01451	0,02279	0,8934
Avenida de Portugal	0,00009	0,00004	0,00438	0,00652	0,87673
Calle Gran Via	0,00138	0,00223	0,03499	0,04723	0,54184
Calle Atocha	0,00048	0,001	0,02345	0,03167	0,60539
Avenida de Oporto	0,00004	0,00005	0,00504	0,00685	0,82256
Avenida del Manzanares (M-30)	0,00385	0,00751	0,06229	0,08667	0,89729
Calle Jose Abascal	0,00105	0,00255	0,0361	0,0505	0,66418
Calle Genova	0,00068	0,00167	0,02771	0,04082	0,87258
Calle Jose Ortega y Gasset	0,00012	0,00024	0,01017	0,01551	0,87123
Avenida Reina Victoria	0,0013	0,0034	0,03486	0,05834	0,90993
Calle Alberto Aguilera	0,0005	0,0009	0,02041	0,03005	0,92615

Продовження таблиці 4.1

1	2	3	4	5	6
Calle Cea Bermudez	0,00069	0,00221	0,03012	0,04703	0,87955
Avenida Menendez Pelayo	0,00071	0,0017	0,02652	0,04118	0,87967
Calle Bravo Murillo	0,0004	0,00088	0,02002	0,02959	0,83251
Avenida del Manzanares (M-30)-2	0,00251	0,01016	0,0665	0,1008	0,90567
Calle Principe de Vergara	0,00023	0,00105	0,01866	0,03238	0,73961
Calle Ronda de Valencia	0,00045	0,00093	0,01978	0,0305	0,84529
Paseo de El Prado	0,00104	0,00222	0,03092	0,04715	0,79614
Calle de Gran Via de San Francisco	0,00092	0,00268	0,03588	0,05173	0,81302
Calle Hortaleza	0,0007	0,00127	0,02573	0,03564	0,46028
Calle San Bernardo	0,00099	0,00187	0,03275	0,04323	0,71805
Calle Alcala	0,00169	0,00404	0,03718	0,06359	0,52432
Calle Mendez Alvaro	0,00092	0,00302	0,0366	0,05499	0,89408
Paseo Infanta Isabel	0,00046	0,00118	0,02257	0,03439	0,8247
Calle Embajadores	0,00034	0,00084	0,01737	0,02894	0,7797
Francos Rodriguez	0,0014	0,00278	0,0323	0,05271	0,87748
Calle Toledo	0,00242	0,00519	0,04847	0,07206	0,80969
Calle Sinesio Delgado	0,0021	0,00663	0,0536	0,08141	0,912
Calle Mayor	0,00022	0,00006	0,00547	0,00767	0,72562

Продовження таблиці 4.1

1	2	3	4	5	6
Paseo de la Castellana-2	0,00089	0,00289	0,03407	0,0538	0,81252
Calle Costa Rica	0,00065	0,00232	0,02922	0,04817	0,89939
Avenida Cardenal Herrera Oria	0,00145	0,0023	0,03202	0,04797	0,92228
Avenida de la Ilustracion (M-30)	0,00125	0,00824	0,05792	0,09078	0,85713
Calle Raimundo Fernandez Villaverde	0,00041	0,00224	0,03051	0,04732	0,82024
Calle Bravo Murillo-2	0,00021	0,00059	0,01556	0,0243	0,87341
Avenida General Peron	0,00036	0,0006	0,01507	0,02457	0,76608
Paseo de Extremadura	0,00221	0,00536	0,05027	0,07322	0,84882
Calle Serrano	0,00146	0,00149	0,02213	0,03856	0,82702
Calle Velazquez	0,0009	0,00163	0,02757	0,04031	0,87282
Avenida de la Albufera	0,00089	0,00233	0,03115	0,04822	0,91176
Calle Alcala-2	0,00183	0,00541	0,05481	0,07357	0,87259
Calle Hermanos Garcia Noblejas	0,00073	0,00252	0,02761	0,05017	0,86665
Avenida de Valladolid	0,00216	0,0064	0,05322	0,07999	0,85899
Calle Lopez de Hoyos	0,00506	0,00411	0,04387	0,06415	0,81251
Avenida Alfonso XIII	0,00234	0,00393	0,04002	0,06266	0,8746
Avenida Brasilia	0,00327	0,00836	0,05908	0,09146	0,75393
Calle de Marcelo Usera	0,00099	0,00711	0,04738	0,08434	0,64232

Кінець таблиці 4.1

1	2	3	4	5	6
Avenida Rafaela Ybarra	0,00093	0,00162	0,02866	0,04026	0,87358
Calle Alcocer	0,00177	0,00436	0,04467	0,06603	0,8799
Avenida Arcentales	0,00025	0,00058	0,01253	0,02398	0,70578
Calle Silvano	0,00175	0,00436	0,04219	0,06606	0,92085
Avenida de Logrono	0,00114	0,00323	0,03608	0,05679	0,86795
Calle San Cipriano	0,0009	0,00258	0,0353	0,05077	0,84474
Calle Camino de Vinateros	0,00024	0,00059	0,01542	0,02429	0,83872

У всіх інших стовпчиках таблиці 4.1 окрім стовпчика навчання приведено значення відповідних показників для тестової вибірки.

На рис. 4.3 приведено приклад отриманих результатів прогнозування на основі однієї з даних LSTM-моделей.

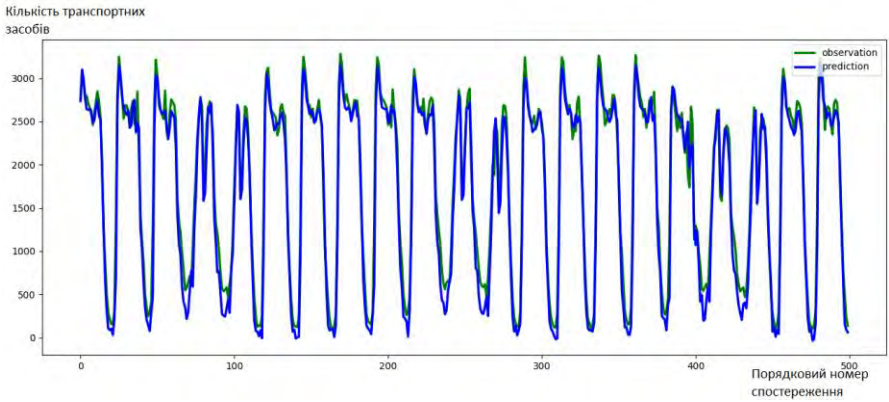


Рисунок 4.3 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 19 лютого по 2 березня 2022 року через годину на основі базової LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин

Приведені на рис. 4.3-4.8 результати стосуються моделі, побудованої для станції Calle Cea Bermudez на основі трафіку за минулі 6 годин за цією ж станцією. На графіках синім кольором показано результат прогнозування моделі, а зеленим – фактичне значення.

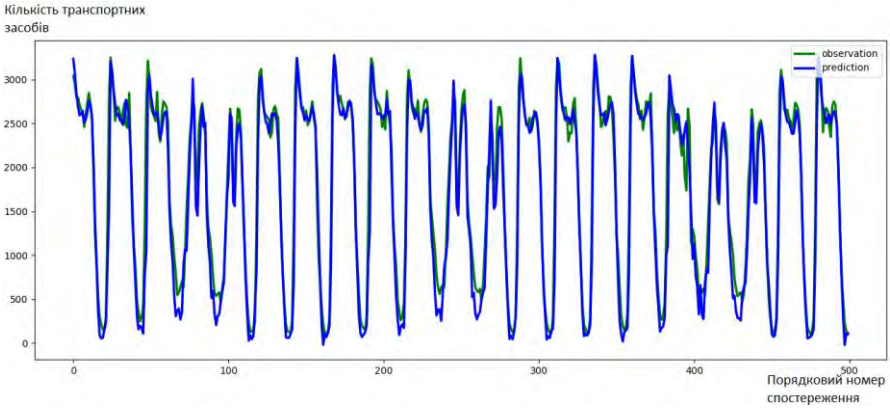


Рисунок 4.4 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 19 лютого по 2 березня 2022 року через 2 години на основі базової LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин

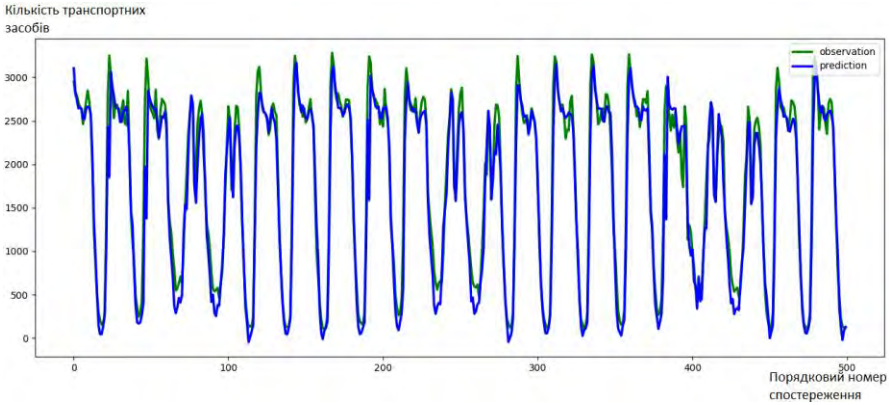


Рисунок 4.5 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 19 лютого по 2 березня 2022 року через 3 години на основі базової LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин

Результати прогнозування на кожному з графіків приведені за відповідним проміжком у майбутньому. Тобто модель дозволяє отримати одразу прогноз трафіку на 1, 2, 3, 4, 5, 6-у годину. За кожним таким прогнозом перші години подані на рис. 4.3, шості – на рис. 4.8.

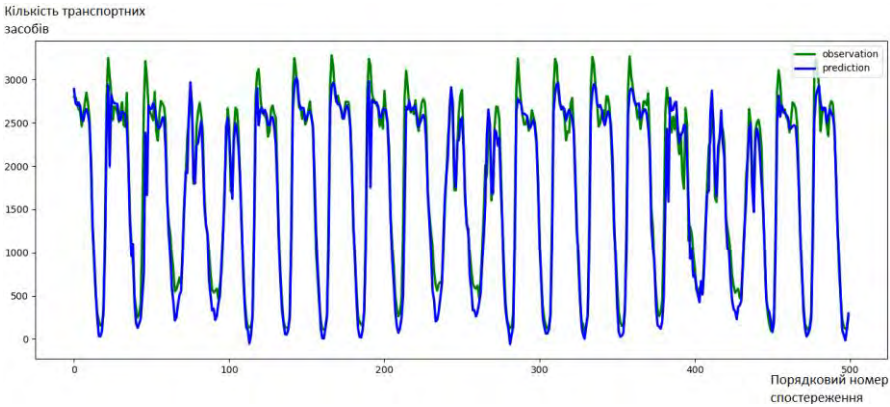


Рисунок 4.6 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 19 лютого по 2 березня 2022 року через 4 години на основі базової LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин

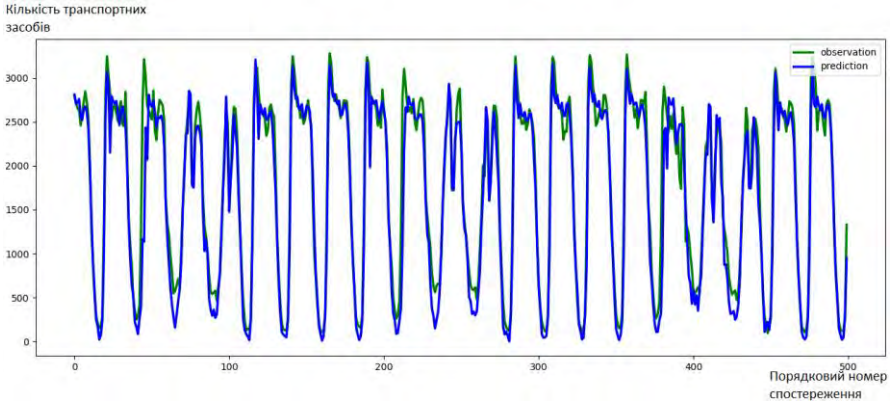


Рисунок 4.7 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 19 лютого по 2 березня 2022 року через 5 години на основі базової LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин

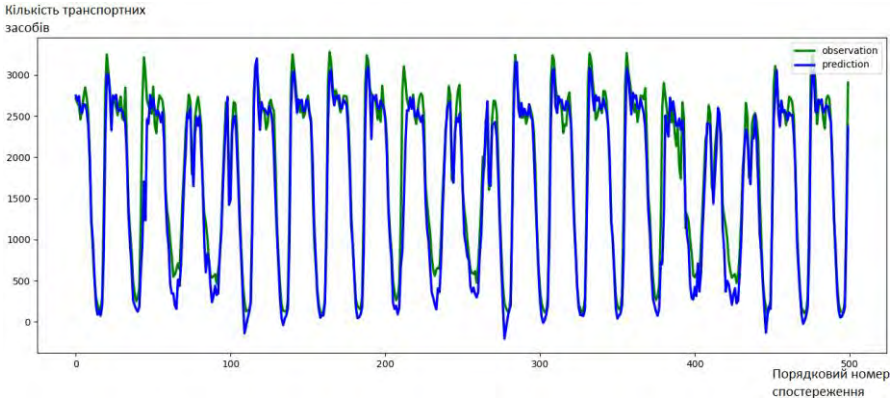


Рисунок 4.8 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 19 лютого по 2 березня 2022 року через 6 годин на основі базової LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин

Аналізуючи отримані результати, можна помітити, що відхилення між очікуваними результатами та отриманими (прогнозованими) зі збільшенням інтервалу прогнозування поступово збільшуються, що зокрема добре помітно при порівнянні результатів на рис. 4.3 та 4.8. Окрім того на рис. 4.8 помітно, що прогнозоване значення в деяких позиціях має помітний лаг порівняно з очікуваним. Відповідно фактично це призводить до запізнення з реагуванням під час прийняття рішень.

У таблиці 4.2 представлено значення аналізованих показників для моделі, результати роботи якої представлено на рис. 4.3-4.8, тобто за станцією Calle Cea Bermudez, в розрізі годин прогнозування. Як помітно з наведених результатів, зі збільшенням періоду прогнозування значення всіх оцінок збільшуються, тобто погіршуються. При цьому для деяких станцій у певні години можливе фактично зупинення такої зміни. Наприклад, у даному випадку значення MSE та RMSE для 4-ої та 5-ої годин є дуже близькими, тому у випадку нормалізації та округлення до 5 знаків значення рівне. Для деяких станцій може траплятися ситуація невеликого покращення відповідних значень на наступну годину, але вже далі значення все одно погіршується, тобто такі випадки все одно не спростовують загальну тенденцію.

Таблиця 4.2 – Результати прогнозування автомобільного трафіку за станцією Calle Cea Bermudez на основі базової LSTM-моделі в розрізі періоду прогнозування

Період прогнозування, годин	MSE	MAE	RMSE
1	0,000931	0,023229	0,030513
2	0,001818	0,027716	0,042637
3	0,002413	0,02948	0,049124
4	0,00258	0,031568	0,050789
5	0,00258	0,032615	0,050789
6	0,002949	0,036085	0,054308

Якщо узагальнювати значення помилки прогнозування за всіма станціями (за створеними базовими LSTM-моделями), то відповідна гістограма показана на рис. 4.9. На ній відображено кількість моделей, значення MSE, для яких потрапляє у відповідний інтервал. Якщо розглядати зміну MSE протягом 1-6 годин прогнозування, то відповідні гістограми показані на рис. 4.10-4.15. На рис. 4.15 при загальній подібності розподілу можна помітити зокрема, що розмір інтервалу збільшився майже у 3 рази порівняно з рис. 4.10.

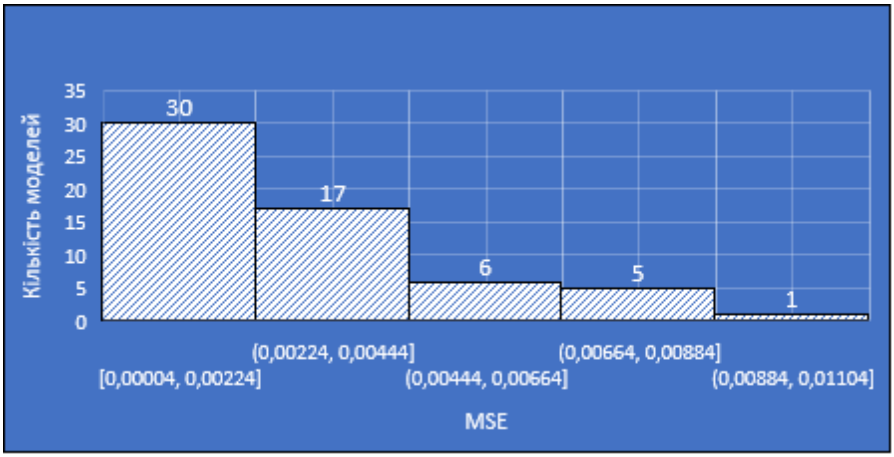


Рисунок 4.9 – Гістограма розподілу результатів прогнозування трафіку базових LSTM-моделей за MSE

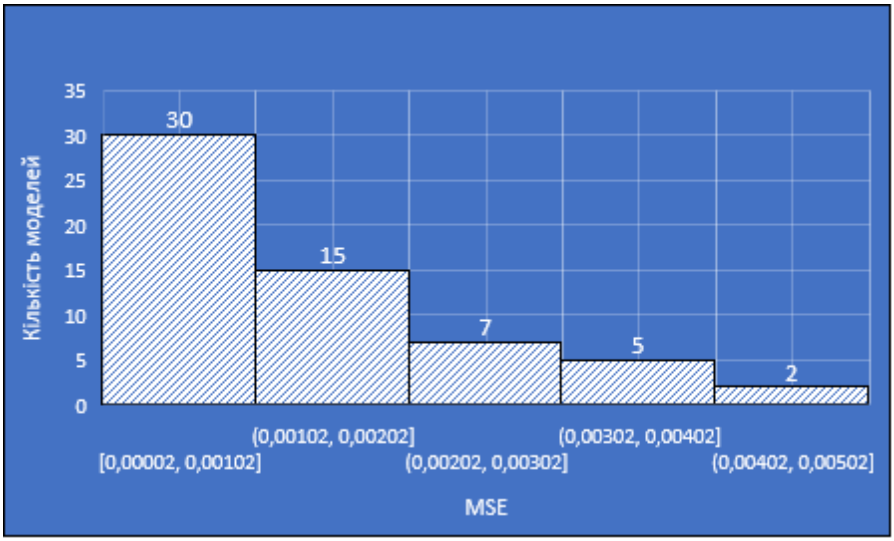


Рисунок 4.10 – Гістограма розподілу результатів прогнозування трафіку на 1 годину вперед базових LSTM-моделей за MSE

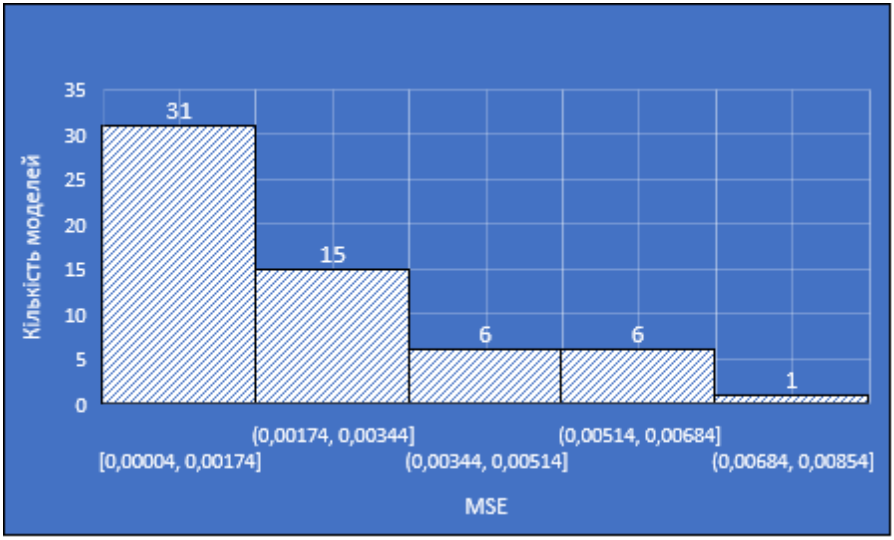


Рисунок 4.11 – Гістограма розподілу результатів прогнозування трафіку на 2 години вперед базових LSTM-моделей за MSE

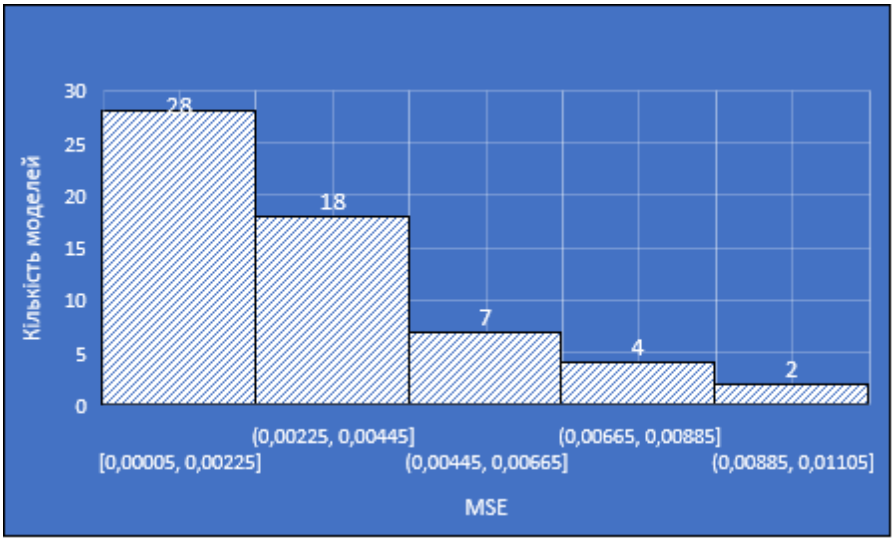


Рисунок 4.12 – Гістограма розподілу результатів прогнозування трафіку на 3 години вперед базових LSTM-моделей за MSE

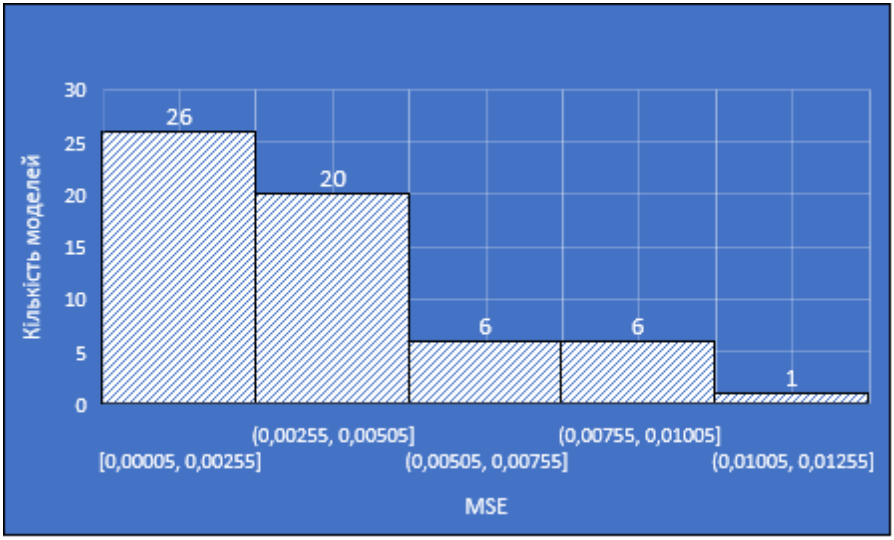


Рисунок 4.13 – Гістограма розподілу результатів прогнозування трафіку на 4 години вперед базових LSTM-моделей за MSE

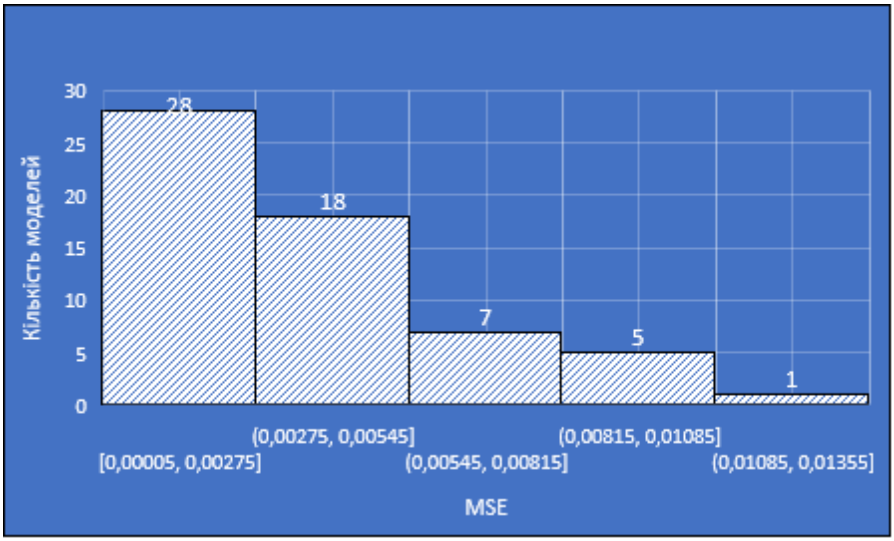


Рисунок 4.14 – Гістограма розподілу результатів прогнозування трафіку на 5 годин вперед базових LSTM-моделей за MSE

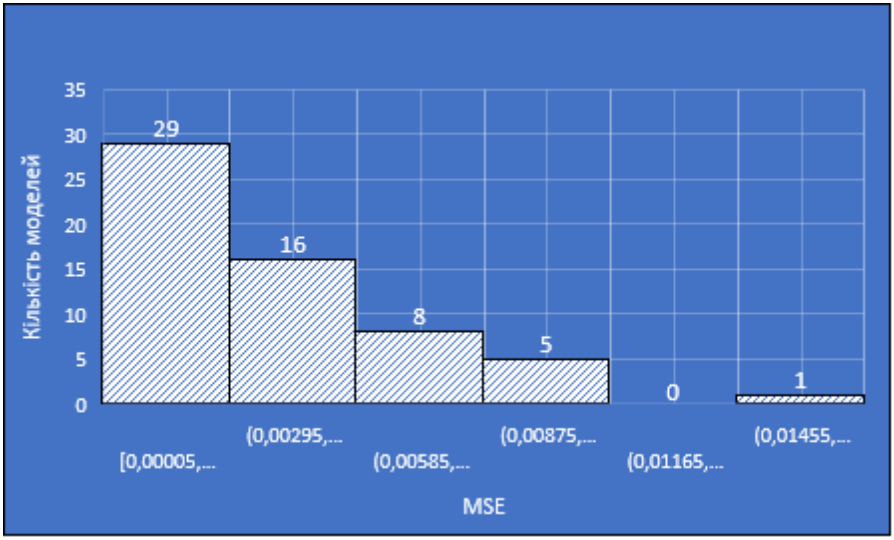


Рисунок 4.15 – Гістограма розподілу результатів прогнозування трафіку на 6 годин вперед базових LSTM-моделей за MSE

Наступними були створені двонаправлені LSTM-моделі (biLSTM) з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку. Тобто вхідні ознаки були такими самими як і в попередній моделі, але внутрішня структура цієї моделі відрізняється.

Особливістю biLSTM-моделей є те, що вони використовують у своїй структурі дві LSTM, пропускаючи таким чином вхідні дані в прямому та зворотному порядках. Це дозволяє витягнути більше інформації про наявні залежності у вхідних даних у процесі навчання.

biLSTM-моделі було використано за наступною структурою:

– вхідний шар, який складається з 6 ознак (значення трафіку за минулі 6 годин);

– перший двонаправлений шар;

– виключення (dropout) у 0,1;

– другий двонаправлений шар;

– повнозв'язний шар з 6 нейронами для формування остаточних результатів.

Отримані результати тестування створених biLSTM-моделей за розподілом значень MSE представлено на рис. 4.16.

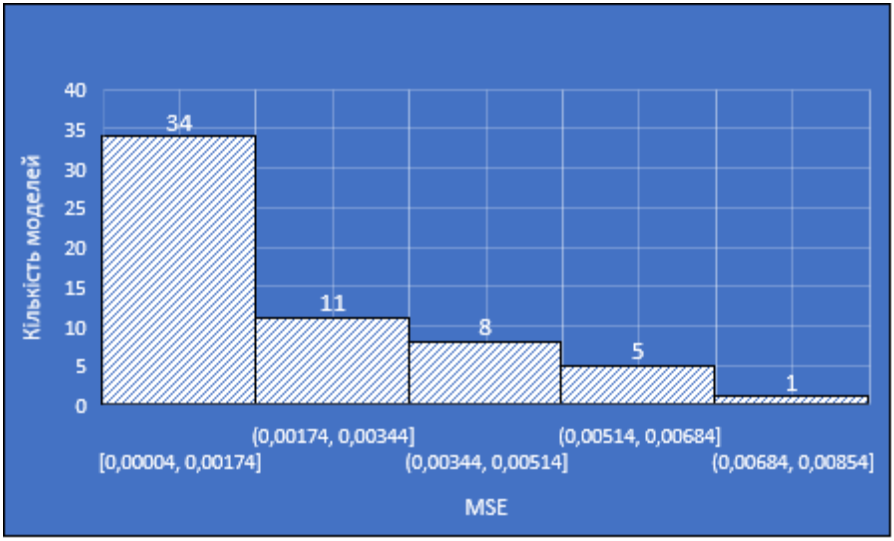


Рисунок 4.16 – Гістограма розподілу результатів прогнозування трафіку на 6 годин вперед biLSTM-моделей за MSE

Наступними були створені LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку, та станціями, які визначені як релевантні для станції прогнозування. Тобто це базові моделі, які мають розширену кількість вхідних ознак. Замість 6 значень для кожного екземпляра подається 6 значень за 3 ознаками: за трафіком самої станції прогнозування та за трафіком інших станцій, які були визначені релевантними даній. Максимальна кількість таких станцій була визначена рівною 2. Тобто на вхід тоді подавалося 3 ознаки з 6 значеннями: всього 18 значень у форматі 3x6.

Відбір ознак, тобто станцій, виконувався на основі застосування ансамблів дерев рішень за допомогою методу Random Forest. Застосування методу Random Forest відбувалось через виділення значень рівня трафіку через годину від поточного часу (тобто з кожного екземпляра перше значення з вихідних значень часового ряду) та набору значень за кожною станцією, яка входить у множину *B*. Серед даного набору виділялись максимально 2 станції, дані яких мали найвищу інформативність для станції прогнозування. Проте якщо інформативність однієї окремої станції або тільки станції прогнозування була вищою для значень рівня трафіку через годину від поточного часу, то міг використовуватись один з таких варіантів. У підсумку було сформовано для кожної станції підмножину релевантних станцій з максимально 2 елементів. Фактично для всіх станцій у результаті було визначено по 2 релевантні станції, а перелік цих станцій приведено в таблиці 4.3.

Таблиця 4.3 – Результати відбору вхідних ознак за станціями вимірювання трафіку на основі методу Random Forest

Станція	Перша вхідна ознака	Друга вхідна ознака
1	2	3
Paseo de la Castellana	Paseo de la Castellana-2	Calle Hermanos Garcia Noblejas
Calle Princesa	Calle Jose Abascal	Calle Mendez Alvaro
Calle Doctor Esquerdo	Avenida Reina Victoria	Calle Hermanos Garcia Noblejas
Paseo de San Franciso de Sales	Avenida Reina Victoria	Calle Principe de Vergara

Продовження таблиці 4.3

1	2	3
Paseo de Santa Maria de la Cabeza	Paseo Infanta Isabel	Calle San Cipriano
Calle Arturo Soria	Avenida Reina Victoria	Avenida del Manzanares (M-30)-2
Avenida de Portugal	Avenida del Manzanares (M-30)	Calle Mendez Alvaro
Calle Gran Via	Calle Atocha	Calle San Cipriano
Calle Atocha	Calle Jose Ortega y Gasset	Calle Ronda de Valencia
Avenida de Oporto	Paseo de San Franciso de Sales	Calle Arturo Soria
Avenida del Manzanares (M-30)	Avenida del Manzanares (M-30)-2	Calle Sinesio Delgado
Calle Jose Abascal	Calle Cea Bermudez	Avenida del Manzanares (M-30)-2
Calle Genova	Calle Alberto Aguilera	Avenida de la Ilustracion (M-30)
Calle Jose Ortega y Gasset	Avenida de Oporto	Calle Genova
Avenida Reina Victoria	Avenida del Manzanares (M-30)-2	Calle Mendez Alvaro
Calle Alberto Aguilera	Calle Genova	Calle Hermanos Garcia Noblejas
Calle Cea Bermudez	Calle Mendez Alvaro	Avenida Rafaela Ybarra
Avenida Menendez Pelayo	Avenida del Manzanares (M-30)-2	Calle Silvano
Calle Bravo Murillo	Calle Arturo Soria	Calle Alberto Aguilera
Avenida del Manzanares (M-30)-2	Avenida del Manzanares (M-30)	Calle Alberto Aguilera
Calle Principe de Vergara	Calle Jose Ortega y Gasset	Calle Hermanos Garcia Noblejas
Calle Ronda de Valencia	Avenida del Manzanares (M-30)-2	Calle Silvano

Продовження таблиці 4.3

1	2	3
Paseo de El Prado	Paseo de la Castellana	Paseo de Extremadura
Calle de Gran Via de San Francisco	Calle Atocha	Calle Raimundo Fernandez Villaverde
Calle Hortaleza	Avenida de Portugal	Avenida Reina Victoria
Calle San Bernardo	Calle Atocha	Calle Serrano
Calle Alcala	Calle Genova	Avenida de la Albufera
Calle Mendez Alvaro	Avenida de Oporto	Calle Hermanos Garcia Noblejas
Paseo Infanta Isabel	Paseo de Santa Maria de la Cabeza	Calle Hermanos Garcia Noblejas
Calle Embajadores	Calle Mendez Alvaro	Calle Alcocer
Francos Rodriguez	Calle Costa Rica	Avenida Cardenal Herrera Oria
Calle Toledo	Calle Cea Bermudez	Calle de Gran Via de San Francisco
Calle Sinesio Delgado	Calle Costa Rica	Avenida de Valladolid
Calle Mayor	Calle Costa Rica	Calle Raimundo Fernandez Villaverde
Paseo de la Castellana-2	Calle Serrano	Calle Hermanos Garcia Noblejas
Calle Costa Rica	Calle Hermanos Garcia Noblejas	Avenida Rafaela Ybarra
Avenida Cardenal Herrera Oria	Avenida Reina Victoria	Calle Silvano
Avenida de la Ilustracion (M-30)	Calle Sinesio Delgado	Calle Costa Rica
Calle Raimundo Fernandez Villaverde	Avenida del Manzanares (M-30)-2	Avenida General Peron
Calle Bravo Murillo-2	Avenida del Manzanares (M-30)	Calle Costa Rica

Кінець таблиці 4.3

1	2	3
Avenida General Peron	Paseo de la Castellana	Calle Hermanos Garcia Noblejas
Paseo de Extremadura	Calle Costa Rica	Calle Camino de Vinateros
Calle Serrano	Calle Hermanos Garcia Noblejas	Calle Silvano
Calle Velazquez	Avenida de la Ilustracion (M-30)	Calle Serrano
Avenida de la Albufera	Avenida Cardenal Herrera Oria	Calle Hermanos Garcia Noblejas
Calle Alcala-2	Calle Mendez Alvaro	Calle Sinesio Delgado
Calle Hermanos Garcia Noblejas	Calle Mendez Alvaro	Avenida Rafaela Ybarra
Avenida de Valladolid	Calle Hermanos Garcia Noblejas	Calle Silvano
Calle Lopez de Hoyos	Calle Bravo Murillo	Avenida General Peron
Avenida Alfonso XIII	Calle Doctor Esquerdo	Calle Hermanos Garcia Noblejas
Avenida Brasilia	Avenida Arcentales	Calle Silvano
Calle de Marcelo Usera	Calle Cea Bermudez	Calle Alcocer
Avenida Rafaela Ybarra	Calle Alcocer	Calle Silvano
Calle Alcocer	Avenida de la Albufera	Calle Hermanos Garcia Noblejas
Avenida Arcentales	Francos Rodriguez	Calle Costa Rica
Calle Silvano	Avenida de la Albufera	Calle Alcocer
Avenida de Logrono	Francos Rodriguez	Calle Silvano
Calle San Cipriano	Calle Costa Rica	Avenida Rafaela Ybarra
Calle Camino de Vinateros	Francos Rodriguez	Calle Costa Rica

Отримані результати тестування створених LSTM-моделей, що складалися з 2 прихованих шарів, на основі вхідних даних з 3 станцій за 6 минулих годин за розподілом значень MSE за всіма моделями представлено на рис. 4.17. Помітним є зменшення інтервалів, тобто фактично похибки за подібного розподілу порівняно з базовими моделями (рис. 4.9), що характерно і для biLSTM-моделей (рис. 4.16). При цьому похибка biLSTM-моделей, як помітно з довжини інтервалів та розподілу, є дещо меншою.

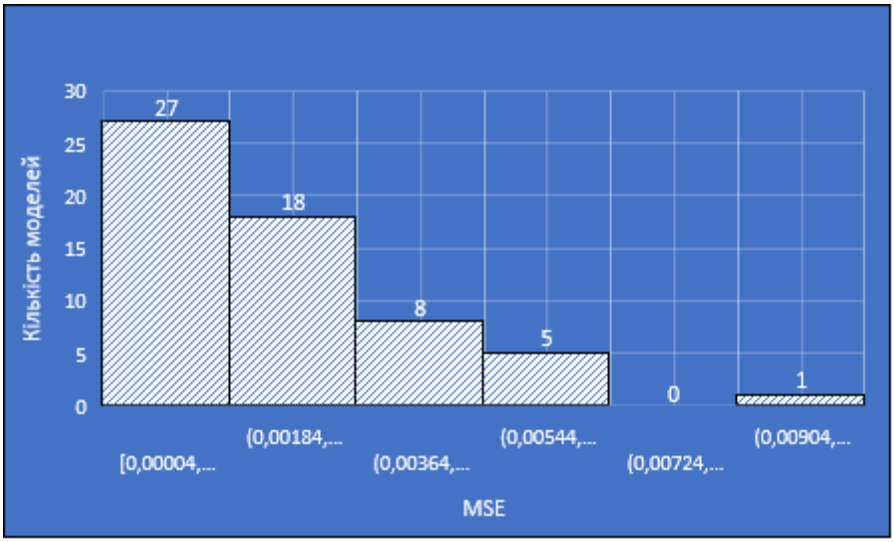


Рисунок 4.17 – Гістограма розподілу результатів прогнозування трафіку на 6 годин вперед LSTM-моделей з вхідними даними з 3 станцій за MSE

Для того щоб дослідити вплив тривалості часового ряду, було створено LSTM-моделі, на вхід яких подавалось послідовність значень, що відповідають трафіку за станцією прогнозування протягом більшої кількості годин у минулому. Для визначення необхідної кількості годин для подачі на вхід моделі було створено графіки, які визначають автокореляцію трафіка за кожною станцією, тобто визначають те, на скільки корелює поточний рівень трафіку за станцією з рівнем трафіку протягом минулих 72 годин за цією ж станцією (рис. 4.18-4.21).

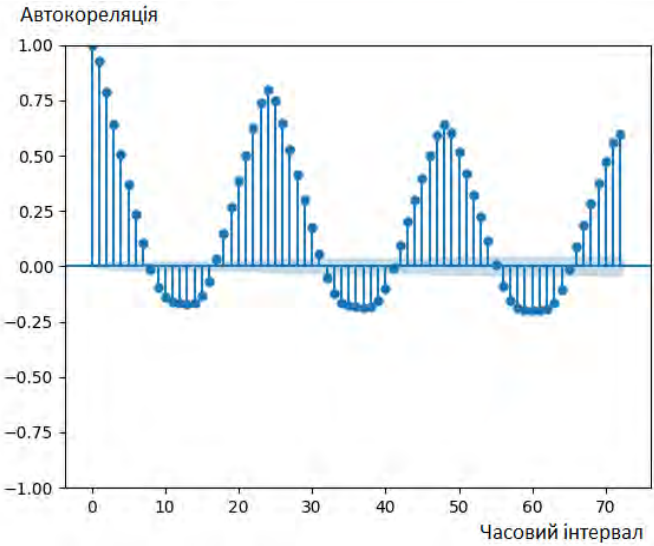


Рисунок 4.18 – Графік автокореляції трафіку за станцією Paseo de la Castellana

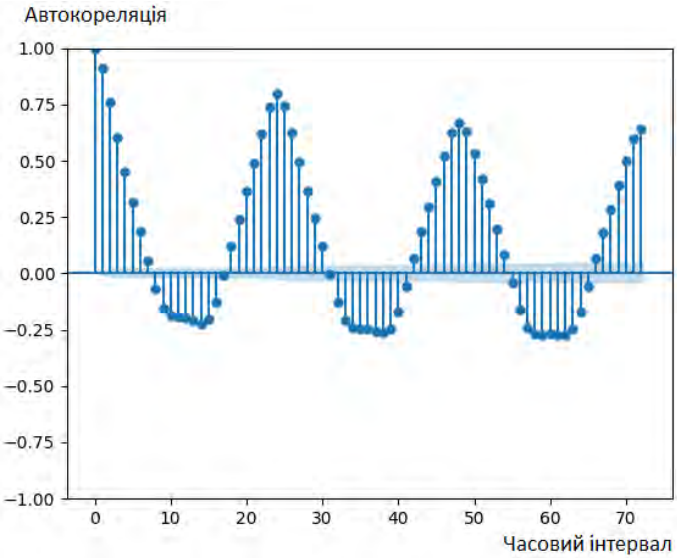


Рисунок 4.19 – Графік автокореляції трафіку за станцією Paseo de San Franciso de Sales

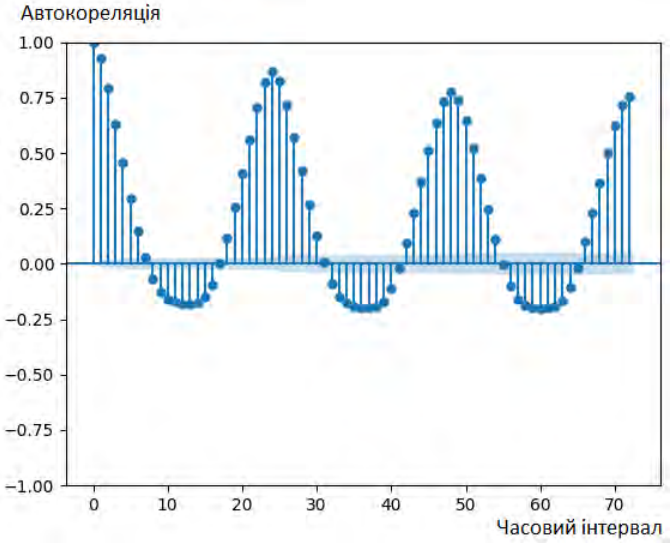


Рисунок 4.20 – Графік автокореляції трафіку за станцією Paseo Infanta Isabel

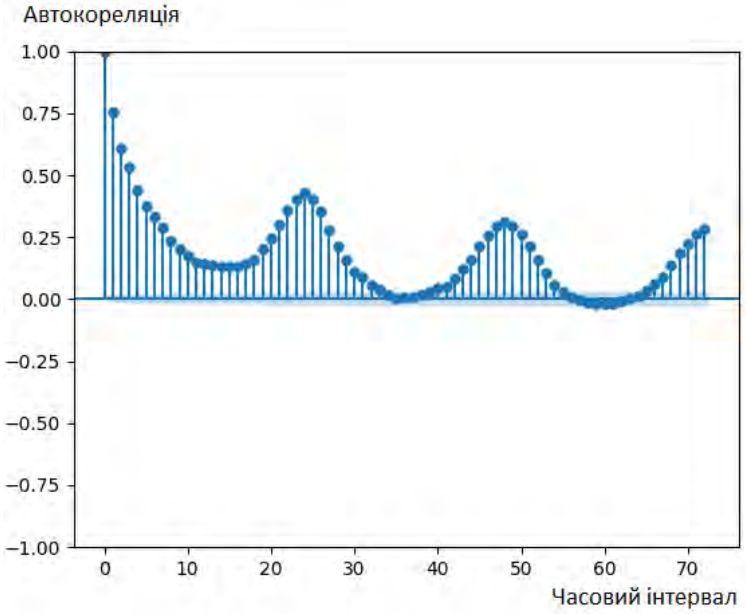


Рисунок 4.21 – Графік автокореляції трафіку за станцією Calle Mayor

Графіки на рис. 4.18.-4.21 представляють приклади деяких станцій зі всього набору з 59 станцій, де відбуваються спостереження. Графіки на рис. 4.18-4.19 загалом є типовими. З них добре видно, як автокореляція знижується протягом 7 годин до нульового рівня, але потім знов підвищується, досягаючи локального піку на 24-у годину, після чого знов знижується. Ці графіки демонструють, що дані за 24 години, тобто за добу мають достатньо високу автокореляцію, а в подальшому знижуються, відповідно дані за 2-у добу вже не є на стільки інформативними. На рис. 4.20 показано приклад більш тривалішої залежності. Для такої станції кількість попередніх годин на вхід моделі можна збільшити. Однак набагато важливішими є випадки окремих станцій, за якими автокореляція драматично знижується вже протягом першої доби (рис. 4.21). У такому випадку витягнути необхідну інформацію про залежність від даних трафіку зі збільшенням інтервалу поданих даних малоймовірно. Тому при створенні результуючих моделей було визначено, що у випадках, коли автокореляція протягом доби спадає до рівня, близького до 0,5, на вхід подається період у 6 годин. Для інших випадків (стандартно за наявними даними) на вхід подається інтервал у 24 попередні години.

Відповідний розподіл отриманої величини для створених моделей з поданими на вхід значеннями трафіку за минулі 24 години та 2 прихованими шарами представлено на рис. 4.22. На даному графіку можна помітити достатньо близькі результати порівняно зі стандартними biLSTM-моделями, тобто моделями побудованими тільки на основі даних за минулі 6 годин, зважаючи зокрема на рівні інтервали. Відбувається близький, але дещо відмінний розподіл між категоріями за величиною MSE.

Оскільки створені та навчені LSTM-моделі з вхідними ознаками на основі даних трафіку за минулі 6 годин за станцією, за якою відбувається прогнозування трафіку, та метеорологічними даними, співвіднесеними з цією станцією за цей же період часу, не призвели до покращення результатів порівняно з базовою LSTM-моделлю, а в багатьох варіаціях відбулось певне погіршення результатів, то результати цих моделей не було приведено в роботі, зважаючи на великий обсяг отриманих результатів та важливість акцентувати увагу на виявлених тенденціях для подальшого практичного застосування.

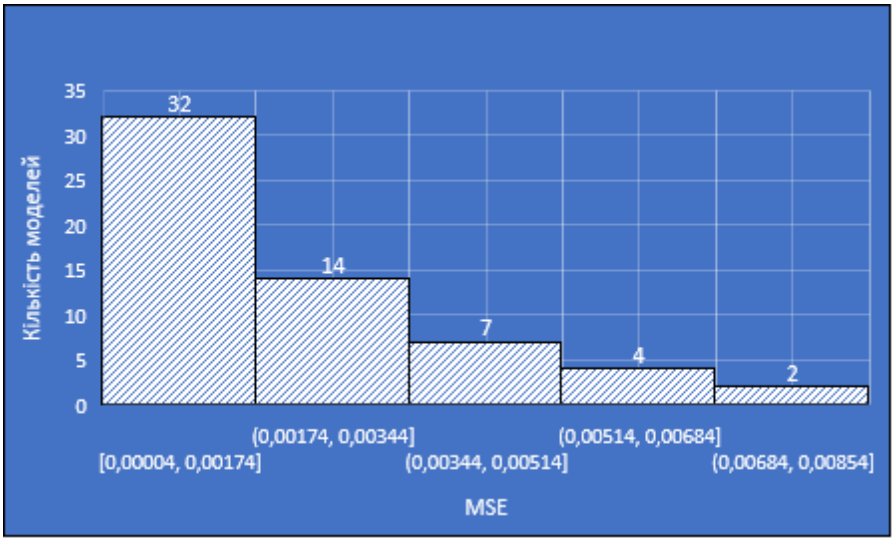


Рисунок 4.22 – Гістограма розподілу результатів прогнозування трафіку на 6 годин вперед LSTM-моделей з вхідними даними за станцією прогнозування за 24 попередні години за MSE

Об’єднуючи рішення, які дозволили отримати певні покращення результатів прогнозування, були biLSTM-моделі з вхідними ознаками на основі даних трафіку за минулу оптимальну (вибрану відповідним чином) кількість годин за станцією, за якою відбувається прогнозування трафіку, та станціями, які визначені як релевантні для станції прогнозування. Тобто в цих моделях об’єднано вхідні ознаки, які призвели до покращення загальних результатів на моделях, за якими фактично було досліджено результати з окремими вхідними ознаками. Вибір періоду вхідних даних (6 або 24 години) виконувався на основі принципу перевірки даних за автокореляцією, що було описано вище. Якщо під кінець першої доби під час пікових значень автокореляція досягає значення, близького до 0,5, але значно його не перевищує, то використовується інтервал у 6 годин. У протилежному випадку використовується інтервал у 24 години.

Результати виконаного тестування для результуючих моделей, побудованих на основі запропонованого в даному підрозділі методу з 2 прихованими biLSTM-шарами за всіма станціями, які були проаналізовані в роботі, приведені в таблиці 4.4.

Таблиця 4.4 – Результати навчання та тестування результуючих biLSTM-моделей з вхідними ознаками на основі даних трафіку за минулі 24 години за станцією, за якою відбувається прогнозування трафіку, та станціями, які визначені як релевантні для станції прогнозування

Станція	Навчання	MSE	MAE	RMSE	R^2
1	2	3	4	5	6
Paseo de la Castellana	0,00029	0,00121	0,02304	0,03478	0,91366
Calle Princesa	0,00062	0,00304	0,03969	0,05511	0,86862
Calle Doctor Esquerdo	0,00062	0,00318	0,03538	0,05642	0,93798
Paseo de San Franciso de Sales	0,00013	0,00028	0,01142	0,01674	0,92854
Paseo de Santa Maria de la Cabeza	0,00025	0,00095	0,02219	0,03075	0,93774
Calle Arturo Soria	0,00006	0,00024	0,01086	0,01555	0,95032
Avenida de Portugal	0,00009	0,00004	0,00436	0,00644	0,87975
Calle Gran Via	0,00036	0,00181	0,03019	0,04254	0,62886
Calle Atocha	0,00017	0,00056	0,01658	0,02356	0,78185
Avenida de Oporto	0,00002	0,00003	0,00426	0,00568	0,87835
Avenida del Manzanares (M-30)	0,0011	0,00497	0,04981	0,07048	0,93223
Calle Jose Abascal	0,0003	0,00109	0,02331	0,03298	0,85686
Calle Genova	0,00019	0,00074	0,01812	0,02725	0,94329
Calle Jose Ortega y Gasset	0,00004	0,00011	0,00766	0,01056	0,94029

Продовження таблиці 4.4

1	2	3	4	5	6
Avenida Reina Victoria	0,00052	0,00173	0,0285	0,04157	0,9542
Calle Alberto Aguilera	0,00019	0,00048	0,01534	0,02187	0,96093
Calle Cea Bermudez	0,00022	0,00085	0,02073	0,02918	0,95364
Avenida Menendez Pelayo	0,0002	0,00062	0,01736	0,02482	0,95632
Calle Bravo Murillo	0,00013	0,00038	0,01404	0,01946	0,92763
Avenida del Manzanares (M-30)-2	0,00074	0,00434	0,05105	0,06586	0,95976
Calle Principe de Vergara	0,00007	0,00042	0,01321	0,02039	0,8967
Calle Ronda de Valencia	0,00017	0,00048	0,01573	0,02196	0,92001
Paseo de El Prado	0,00031	0,00153	0,02657	0,03912	0,8599
Calle de Gran Via de San Francisco	0,00029	0,00233	0,03432	0,04824	0,83752
Calle Hortaleza	0,00027	0,00101	0,02299	0,0318	0,57113
Calle San Bernardo	0,00042	0,00113	0,02502	0,03354	0,83073
Calle Alcala	0,00061	0,00349	0,03211	0,05911	0,58887
Calle Mendez Alvaro	0,0004	0,00173	0,03009	0,04162	0,93931
Paseo Infanta Isabel	0,00018	0,00061	0,01838	0,02478	0,90909
Calle Embajadores	0,00011	0,00026	0,01098	0,01607	0,93215

Продовження таблиці 4.4

1	2	3	4	5	6
Francos Rodriguez	0,0004	0,0011	0,0217	0,0331	0,95163
Calle Toledo	0,00091	0,00284	0,03681	0,05327	0,89604
Calle Sinesio Delgado	0,00055	0,00348	0,0419	0,05902	0,95376
Calle Mayor	0,00013	0,00005	0,00527	0,00717	0,76075
Paseo de la Castellana-2	0,00027	0,00137	0,02571	0,03704	0,91111
Calle Costa Rica	0,00023	0,00082	0,0204	0,02864	0,96447
Avenida Cardenal Herrera Oria	0,00052	0,00157	0,02676	0,03965	0,94687
Avenida de la Ilustracion (M-30)	0,00042	0,00526	0,04489	0,07251	0,90892
Calle Raimundo Fernandez Villaverde	0,00016	0,00144	0,02469	0,03794	0,8842
Calle Bravo Murillo-2	0,00007	0,00026	0,01072	0,01602	0,94508
Avenida General Peron	0,00011	0,00035	0,01138	0,01865	0,86467
Paseo de Extremadura	0,00064	0,00316	0,03875	0,05617	0,91113
Calle Serrano	0,00054	0,00114	0,01819	0,03374	0,86764
Calle Velazquez	0,00026	0,00098	0,02179	0,03131	0,92323
Avenida de la Albufera	0,00031	0,00121	0,02352	0,03477	0,95417
Calle Alcala-2	0,00067	0,00276	0,04026	0,0525	0,93512
Calle Hermanos Garcia Noblejas	0,00021	0,00133	0,02096	0,03642	0,92972

Кінець таблиці 4.4

1	2	3	4	5	6
Avenida de Valladolid	0,00077	0,00368	0,04043	0,06064	0,9191
Calle Lopez de Hoyos	0,00182	0,00249	0,03487	0,04994	0,88648
Avenida Alfonso XIII	0,00087	0,00291	0,03758	0,05392	0,9068
Avenida Brasilia	0,00109	0,00489	0,04746	0,06996	0,8559
Calle de Marcelo Usera	0,00036	0,00473	0,03677	0,06876	0,76216
Avenida Rafaela Ybarra	0,00034	0,00076	0,02137	0,02757	0,94082
Calle Alcocer	0,00076	0,00213	0,03353	0,0462	0,94125
Avenida Arcentales	0,00011	0,00047	0,01019	0,0216	0,76143
Calle Silvano	0,00057	0,00204	0,0324	0,04519	0,96296
Avenida de Logrono	0,00036	0,00142	0,027	0,03764	0,94191
Calle San Cipriano	0,00035	0,00165	0,03003	0,04066	0,90039
Calle Camino de Vinateros	0,00007	0,00035	0,01297	0,0186	0,90508

Для порівняння фактично отриманих значень трафіку було створено ряд графіків, які відображають результати прогнозування трафіку за станцією Calle Cea Bermudez на 1 годину (рис. 4.23), 2 години (рис. 4.24), 3 години (рис. 4.25), 4 години (рис. 4.26), 5 годин (рис. 4.27), 6 годин (рис. 4.28) вперед. Інтервал виведених дат дещо відрізняється від LSTM-моделей, оскільки графіки формувалися програмно на всьому проміжку, при цьому інтервал годин дещо відрізнявся, зокрема через те, що для результуючих моделей на вхід подавалися значення за 24 години, а не 6, відповідно це переносило перший прогноз в часі. Під час досліджень враховувались відповідні рівноцінні проміжки часу.

Отримані результати, порівняно з базовими LSTM-моделями, близькі відносно фактичних даних з тестувальної вибірки, при чому як за амплітудою значень, конкретними значеннями, так і за відсутністю затримки, яка з часом ставала вираженою у LSTM-моделей.

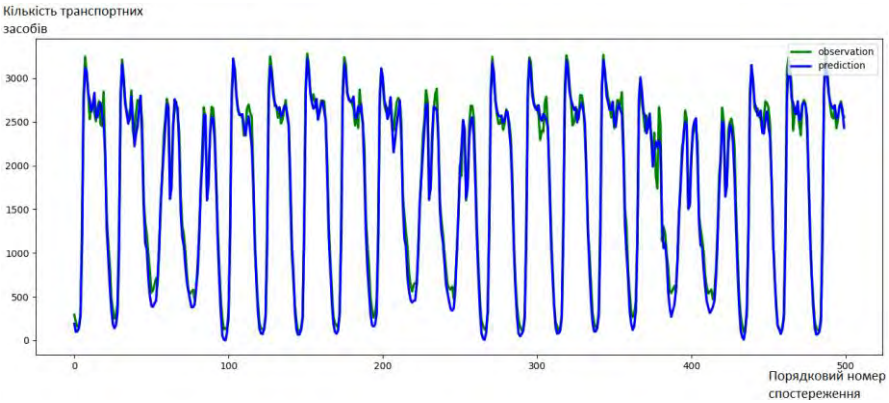


Рисунок 4.23 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 10 лютого по 2 березня 2022 року через годину на основі результуючої LSTM-моделі з вхідними ознаками на основі даних трафіку цієї та релевантних станцій за минулі 24 години

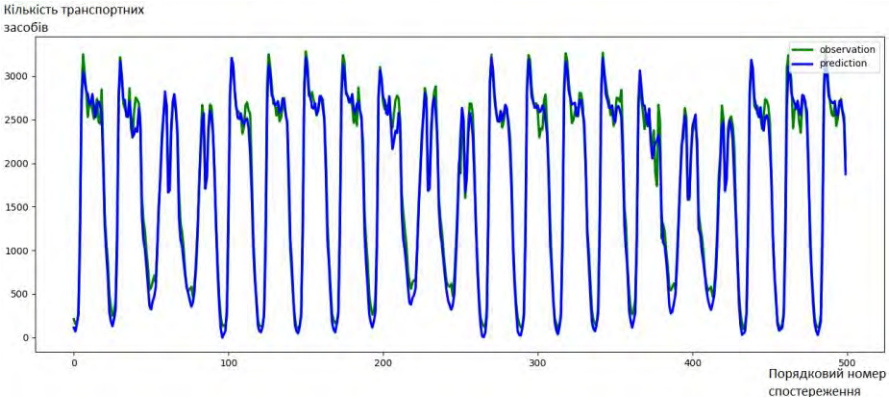


Рисунок 4.24 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 10 лютого по 2 березня 2022 року через 2 години на основі результуючої LSTM-моделі з вхідними ознаками на основі даних трафіку цієї та релевантних станцій за минулі 24 години

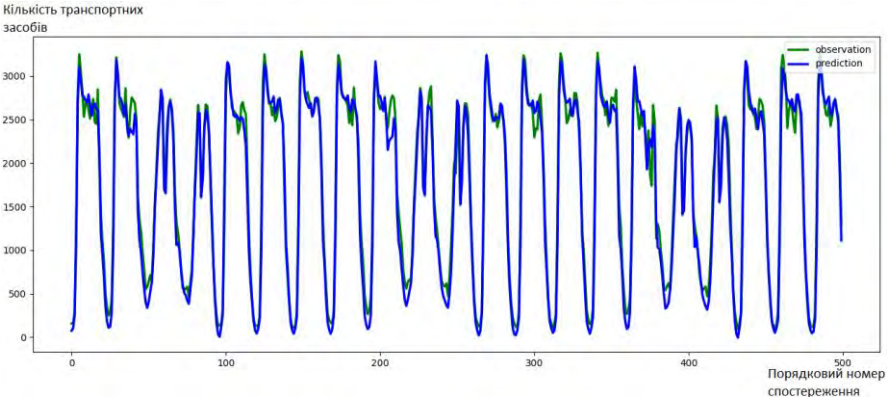


Рисунок 4.25 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 10 лютого по 2 березня 2022 року через 3 години на основі результуючої LSTM-моделі з вхідними ознаками на основі даних трафіку цієї та релевантних станцій за минулі 24 години

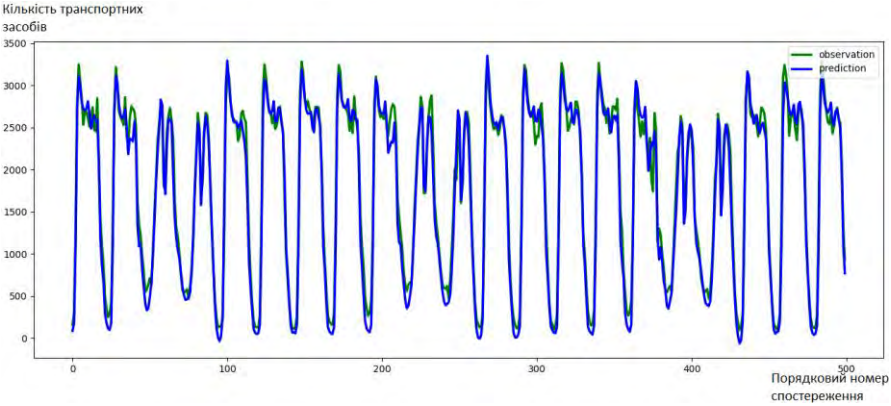


Рисунок 4.26 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 10 лютого по 2 березня 2022 року через 4 години на основі результуючої LSTM-моделі з вхідними ознаками на основі даних трафіку цієї та релевантних станцій за минулі 24 години

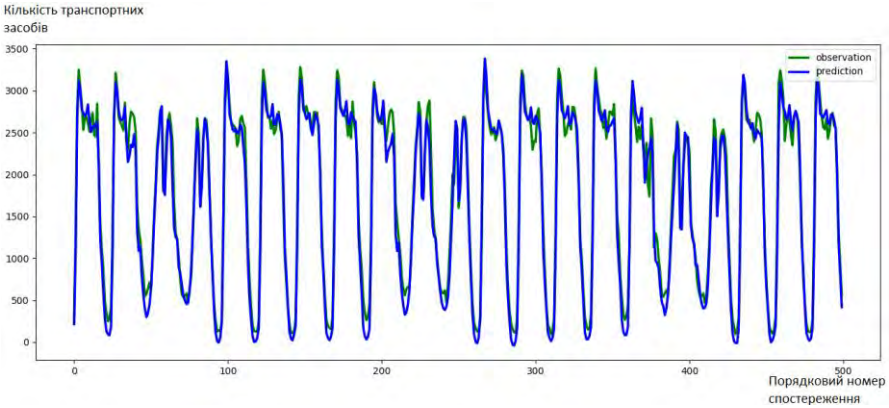


Рисунок 4.27 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 10 лютого по 2 березня 2022 року через 5 годин на основі результуючої LSTM-моделі з вхідними ознаками на основі даних трафіку цієї та релевантних станцій за минулі 24 години

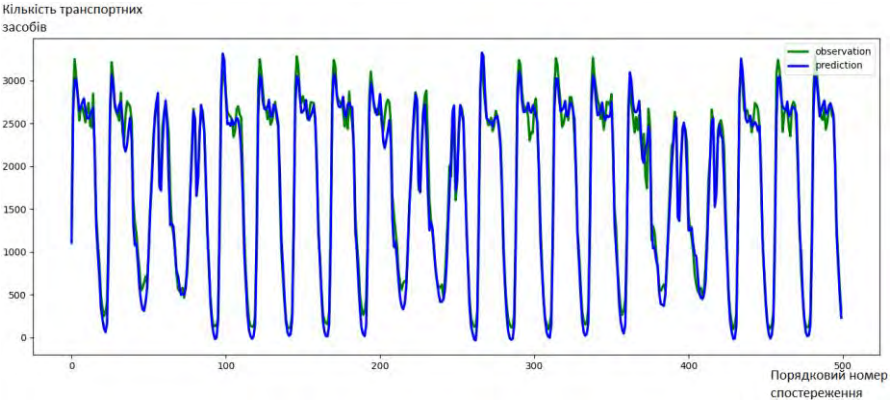


Рисунок 4.28 – Прогнозування кількості транспортних засобів за станцією Calle Cea Bermudez з 10 лютого по 2 березня 2022 року через 6 годин на основі результуючої LSTM-моделі з вхідними ознаками на основі даних трафіку цієї та релевантних станцій за минулі 24 години

Аналогічний розглянутим вище моделям розподіл рівня помилки MSE представлено для результуючих запропонованих моделей на рис. 4.29. Інтервали помилок порівняно з попередніми представленими моделями зменшились.

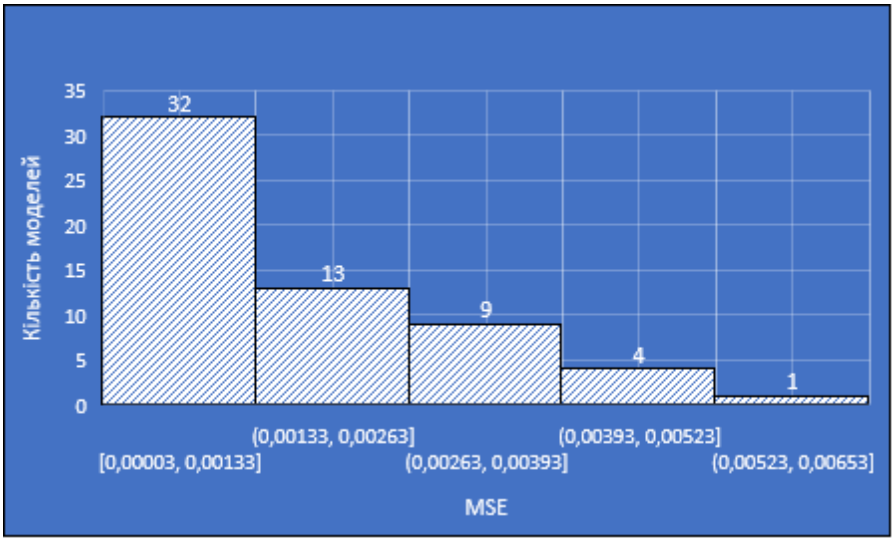


Рисунок 4.29 – Гістограма розподілу результатів прогнозування трафіку на 6 годин вперед biLSTM-моделей з вхідними даними за станцією прогнозування та 2 релевантними станціями за оптимальну кількість попередніх годин за MSE

Результати експериментального дослідження на основі тестувальної вибірки за всіма описаними моделями зведені до таблиці 4.5 на основі розрахованих значень показників MSE, MAE, RMSE, R^2.

Таблиця 4.5 – Середні результати прогнозування автомобільного трафіку на наступні 6 годин на основі різних моделей

Модель прогнозування	MSE	MAE	RMSE	R^2
LSTM	0,002908	0,032192	0,049139	0,818415
LSTM з 3 станціями	0,002356	0,028937	0,043898	0,851874
LSTM за 24 години	0,00215	0,028496	0,04205	0,854499
biLSTM	0,002125	0,027964	0,041744	0,863663
Запропонована модель	0,001631	0,024942	0,036556	0,891

Отримані середні результати демонструють значне покращення результатів прогнозування автомобільного трафіку за допомогою підсумкових запропонованих моделей, які враховують і релевантні станції, і використовують оптимізовану тривалість вхідних даних

(фактичну збільшену кількість годин) при аналізі середнього рівня на 59 станціях. Покращення було отримано одразу за всіма показниками, включаючи і MSE, і MAE, і RMSE, і R^2. Тож загалом можна стверджувати про те, що і точність отриманих результатів була вищою порівняно зі всіма моделями, і інформативність, що в підсумку вказує на практичну придатність методу, який дозволив створити дані моделі як у розрізі подальшого застосування отриманих результатів для побудови фреймворку прийняття рішень під час медичного діагностування, так і щодо обґрунтованості цих рішень.

Запропоновані підсумкові моделі дозволили зменшити рівень середньоквадратичної похибки на 43,91 % порівняно з базовою LSTM-моделлю та на 22,52 % за середньою абсолютною похибкою. Інформативність моделей в середньому збільшилась на 8,87 %.

Дані моделі ґрунтуються на поєднанні рішень, кожне з яких призвело до покращення відносно базових LSTM-моделей. З цих рішень найкращим відносно індивідуальних результатів виявилось створення двонаправленої LSTM-моделі, яке самостійно дозволило зменшити середньоквадратичну похибку на 26,92 %, а середню абсолютну похибку – на 13,13 %. У свою чергу запропоновані в підсумку моделі відносно двонаправлених LSTM-моделей зменшили середньоквадратичну похибку на 30,29 %, а середню абсолютну – на 12,12 %.

У підсумку слід зазначити, що створені моделі відрізняються оптимальною кількістю вхідних ознак порівняно з відомими, відповідно за рахунок спрощеної структури потребують менше даних, до того ж не потребують використання цілісної структури станцій у місті на всіх автомобільних дорогах. Тож це дозволяє використовувати дані моделі в умовах обмеження ресурсів та обмеженої доступності даних про трафік у місті. При цьому забезпечується достатньо висока інформативність створених моделей з рівнем точності прогнозування, який значно перевищує зокрема створення базової LSTM-моделі з оптимальною структурою. Створені моделі дозволяють прогнозувати автомобільний трафік на станції спостереження у місті на наступні 6 годин за минулими 24 або в окремих випадках 6 годинами.

4.3 Дослідження математичних моделей прогнозування рівня забрудненості атмосферного повітря для побудови фреймворку прийняття рішень для медичного діагностування

Медичне діагностування полягає у встановленні діагнозу пацієнта, який стає основою для прийняття рішень, які приймає лікар відносно подальшого лікування пацієнта. Тобто від результатів медичного діагностування залежить послідовність наступних рішень. Однак на подальші рішення також впливають наступні обстеження, стан навколишнього середовища, в якому живе людина. Усе це призводить до коригування рішень на наступних етапах, але до того ж впливає на визначення і застосування конкретних процедур, які рекомендовано пацієнту. Наприклад, при виявленні певних хвороб може бути рекомендовано уникати певних ситуацій, манери поведінки тощо.

Частина цих рекомендацій пов’язана з персональною поведінкою, на яку впливає тільки сам пацієнт і відповідно є зоною його виключної відповідальності, але інша частина пов’язана з впливом навколишнього середовища, тому для вчасного і адекватного реагування на такі ситуації і застосування сформульованих рекомендацій потрібно у процесі діагностування і подальшого моніторингу використовувати відповідні моделі прогнозування.

Для постановки діагнозу необхідно сформувати неоднорідний набір даних, що характеризує ситуацію, що спостерігається. З одного боку, ці дані пов'язані з суб'єктивною інформацією про пацієнта, яка визначається під час опитування та різних видів обстеження, а з іншого боку, пов'язані з середовищем проживання пацієнта.

Рівень забрудненості атмосферного повітря є одним з основних показників, що характеризують таке середовище. Залежно від умов навколишнього середовища можна спланувати специфікацію медичного обстеження, результати якої використовуються при прийнятті рішень при постановці діагнозу, та визначити специфіку виконання рішень, прийнятих за результатами діагностики.

Весь набір рішень, прийнятих під час медичної діагностики, формує структуру прийняття рішень, яка складається з наступних етапів:

– прийняття рішень щодо уточнення планового медичного огляду;

– прийняття рішень щодо вибору методів діагностики пацієнта;

– прийняття рішень щодо визначення стану хворого;

– прийняття рішень щодо подальшого лікування хворого та рекомендацій щодо поведінки в певних ситуаціях у майбутньому [13].

Ця група рішень вимагає, з одного боку, накопичення історичних даних про забруднення повітря, тобто показників забруднення повітря певними речовинами, зібраних на відповідних станціях, а з іншого боку, прогнозування рівня забруднення повітря на цих станціях на майбутнє.

Відповідно одними з моделей, які використовуються для отримання даних рішень, є множина моделей прогнозування забрудненості атмосферного повітря. Одночасно дана множина моделей для якісної власної роботи може потребувати даних, які є результатом роботи інших моделей. Відповідно застосування єдиного фреймворку і інтеграція його принципів до створення програмного забезпечення може дозволити вирішити дану проблему системно і сприяти подальшим крокам зв'язування інших проблем в єдину взаємодіючу систему. Саме тому в даному підрозділі увагу приділено моделям прогнозування забрудненості атмосферного повітря.

Задачу прогнозування рівня забрудненості атмосферного повітря слід формулювати як визначення функціональної залежності між рівнем концентрації забруднювача атмосферного повітря p на деякій станції A протягом кожної з H^F годин у майбутньому на основі даних про концентрацію цього забруднювача протягом попередніх H^F годин за цією станцією A, даних про концентрацію інших забруднювачів з множини R^S за станціями з множини I^S, даних про прогнозований автомобільний трафік протягом H^F наступних годин за станціями з підмножини K^S та на основі даних за іншими показниками з множини C:

$$pol_{p,A}^{t+h} = g\begin{pmatrix} pol_{p,A}^{t}, pol_{p,A}^{t-1}, \dots, pol_{p,A}^{t-H^F}, \\ pol_{r,k}^{t}, pol_{r,k}^{t-1}, \dots, pol_{r,k}^{t-H^F}, \dots, \\ tr_b^{t+1}, tr_b^{t+2}, \dots, tr_b^{t+H^W}, \\ v_c^{t}, v_c^{t-1}, \dots, v_c^{t-H^F} \end{pmatrix}, \tag{4.2}$$

$$h = \overline{1, H^F}, r \in R^S, b \in B^C, R^S \subseteq R, B^C \subseteq B, \qquad c \in C, k \in K^S, K^S \subseteq$$

де $pol_{p,A}^{t+h}$ – рівень забрудненості атмосферного повітря

забруднювачем p на станції A протягом $(t+h)$-ої години (з $t+h$ годин 0 хвилин до $t+h$ годин 59 хвилин) у майбутньому;

$pol_{p,A}^{t}, pol_{p,A}^{t-1}, \dots, pol_{p,A}^{t-H^F}$ – забрудненості атмосферного повітря забруднювачем p на станції A протягом t-ої, $(t-1)$-ої, $(t-H^F)$-ої годин у минулому;

K – множина станцій, за якими здійснюється вимірювання рівня забрудненості атмосферного повітря різними забруднювачами;

K^S – підмножина станцій, дані яких є релевантними для прогнозування рівня забрудненості атмосферного повітря для станції A за забруднювачем p;

R^S – підмножина забруднювачів, рівень яких є релевантним для прогнозування рівня забрудненості атмосферного повітря для станції A за забруднювачем p;

R – множина забруднювачів, які розглядаються;

B^C – підмножина станцій вимірювання автомобільного трафіку, яку було обрано з множини B як релевантну для прогнозування рівня забрудненості атмосферного повітря для станції A за забруднювачем p;

C – множина показників, які здійснюють вплив на значення рівня забрудненості атмосферного повітря для станції A за забруднювачем p протягом кожної з $(t+h)$-х годин;

$v_c^t, v_c^{t-1}, \dots, v_c^{t-H^F}$ – значення показника c з множини C протягом t -ої, $(t-1)$-ої, $(t-H^F)$-ої годин у минулому;

t – номер поточної години, тобто момент часу, в який виконується прогнозування.

Форма функціональної залежності g у задачі (4.2) повинна бути досліджена на основі методів машинного навчання. У якості забруднювача, для якого в даній роботі побудовано відповідні моделі, розглянутий діоксид азоту. Такий вибір було зроблено на основі наступних аргументів.

На даний момент для загального оцінювання якості повітря використовують індекс якості повітря. Дана оцінка дозволяє отримати на виході категоріальне значення, що використовується під час моніторингу та прийняття рішень в різних галузях і загалом характеризує стан атмосферного повітря. Індекс якості повітря визначається з врахуванням таких забруднюючих речовин:

– озон (O_3);

– забруднення частинками ($PM_{2.5}$ та PM_{10});
– моноксид вуглецю (CO);
– діоксид сірки (SO_2);
– діоксид азоту (NO_2).

Відповідно слід вказати на те, що точне прогнозування концентрації даних забруднювачів з множини R є критичним також для точного прогнозування індексу якості повітря.

При дослідженні функціональної залежності g потрібно враховувати, що фактори, які визначені як множина C, насправді представляють собою певну сукупність факторів. Часто при знаходженні функціональної залежності g дані фактори не конкретизуються. Однак, зрозуміло, що на рівень забруднення повітря різними речовинами впливає не тільки рівень забруднення в попередній період або періоди, адже кількість відповідної речовини в повітрі не є статичною – вміст збільшується або зменшується з часом, що докладно буде проаналізовано на наявних даних нижче. Відповідно врахування джерела забруднення більш точно є тим фактором прийняття рішень, який потенційно може призвести до отримання більш точних результатів прогнозування концентрації забруднювачів у атмосферному повітрі в майбутньому.

Розглядаючи потенційні джерела забруднення в світлі описаного фреймворку прийняття рішень варто зазначити, що врахування факторів, які є джерелами забруднення має відбуватися не обов'язково через внесення додаткових факторів у створювану модель. Це зокрема пояснюється описаною в фреймворку логікою, тобто ідеєю, що ці дані вже можуть бути наявними для прийняття рішень, адже вони є результатом застосування відповідних окремих моделей, які часто є достатньо якісними щодо вирішення конкретної задачі, яка фактично є підзадачею більш ширшої задачі прийняття рішень.

Відповідно при розв'язанні задач прийняття рішень стосовно медичного діагностування або створення медичної системи загалом чи системи медичного моніторингу важливим є врахування результатів розв'язання задачі прогнозування забрудненості атмосферного повітря певними речовинами. Ця інформація може використовуватися:

– як самостійна для інформування пацієнта;

– для прийняття рішень стосовно рекомендацій пацієнту відносно його поточного стану здоров'я або наявних загальних хронічних захворювань, коли пацієнту може бути рекомендовано переміститися в інше місце, місто або місцевість, де умови навколишнього середовища будуть більш сприятливими для нього;

– для більш детального аналізу стану здоров'я пацієнта на основі результатів моніторингу навколишнього середовища, в якому знаходиться людина.

При цьому слід розуміти, що за рекомендаціями Всесвітньої організації охорони здоров'я згубно діють на організм людини не тільки випадки довготривалого перевищення норм концентрації забруднювачів у атмосферному повітрі, але і короткотривалого. Для обох цих варіантів для кожного забруднювача встановлено відповідні порогові значення.

У свою чергу прогнозування забрудненості атмосферного повітря, як було описано вище, потребує врахування факторів, інформація про які може бути отримана через розв'язання інших пов'язаних задач. Наприклад, задача прогнозування трафіку в певних точках міста може призводити до розв'язання задачі прогнозування забрудненості атмосферного повітря більш точно. Саме на ній і буде зроблено акцент в даному підрозділі. Це не обумовлює використання інформації тільки про одне окреме джерело забруднення, але це визначає, що результати дослідження зв'язку між розв'язанням цих задач у сукупності можуть бути корисними для розв'язання задачі прогнозування забрудненості атмосферного повітря як прикладу застосування описаного фреймворку прийняття рішень для медичного діагностування.

На даний момент існує ряд робіт [13]-[19], в яких виконується прогнозування рівня забрудненості атмосферного повітря загалом. Частина з них стосується діоксиду азоту зокрема або виконує прогнозування тільки для даного забруднювача. Слід також додати, що частина робіт виконує прогнозування не конкретно концентрації забруднювача, а визначення категорії індексу якості повітря. Однак, оскільки в даному підрозділі прогнозується саме концентрація, то такі роботи не приводяться. У цих роботах, так само як це було розглянуто для проблеми прогнозування автомобільного трафіку, виділяється напрямок використання даних про взаємне розміщення станцій на основі застосування згорткових нейронних мереж та інших рішень,

що враховують графову структуру. Проблеми цього підходу були визначені у попередньому підрозділі і залишаються актуальними також і для поточної задачі. В умовах недостатньої кількості даних, коли в місті діє невелика кількість станцій, не накопичено тривалого періоду спостереження (достатнього для підтримки складної структури з великою кількістю параметрів), такий підхід не є актуальним. Він зокрема розглядається в роботі [15], де згорткові LSTM використовуються разом з графовими згортковими нейронними мережами для прогнозування рівня забрудненості атмосферного повітря частками PM2.5. Прогнозування PM10 здійснюється в роботі [16] на основі гібридної моделі, що базується на LSTM та згорткових нейронних мережах.

Ряд інших робіт направлений на застосування моделей на основі глибокого навчання для прогнозування забрудненості атмосферного повітря, що зокрема стосується роботи [17], де такі архітектури були вивчені стосовно прогнозування діоксиду азоту. У роботі [18] виконано дослідження застосування глибоких моделей машинного навчання на основі LSTM та автокодувальників для прогнозування концентрації часток PM2.5 та PM10 у атмосферному повітрі.

У роботі [19] виконується дослідження прогнозування забрудненості атмосферного повітря саме діоксидом азоту на основі двонаправлених згорткових LSTM, тобто використовується гібридна модель. У якості вхідних даних моделі використовуються дані про концентрацію діоксиду азоту в минулий період, метеорологічні дані (швидкість вітру, атмосферний тиск) та дані трафіку (інтенсивність, зайнятість), при цьому через різні причини, зокрема і відсутність даних у певний період, було виключено показники трафіку за минулий період часу (середня швидкість руху транспорту, завантаженість, сонячне випромінювання та опади). Прогнозування здійснювалось на наступні 6 годин за даними попередніх 6 годин.

У подальшому в даному підрозділі розглянуто створення моделей прогнозування рівня забрудненості атмосферного повітря, які загалом є частиною описаного вище фреймворку прийняття рішень для медичного діагностування. Для цього було використано відповідні вибірки даних, що характеризують актуальний стан у місті Мадрид (Іспанія), що розміщені у вільному доступі Порталі відкритих даних Мадридської міської ради [10]:

– вибірка даних стосовно якості повітря в місті Мадрид, що включає погодинні дані з 1 січня 2001 року до 30 вересня 2022 року включно [20];

– вибірка даних стосовно якості повітря в місті Мадрид, що включає щоденні дані з 1 січня 2001 року до 30 вересня 2022 року включно [21];

– вибірка даних стосовно погодних умов у місті Мадрид, що включає погодинні дані з 1 січня 2019 року до 30 вересня 2022 року [12].

Для уточнення цих даних використовувалась також інформація про станції вимірювання рівня забрудненості атмосферного повітря [22] та про метеостанції [23].

Вибірка даних стосовно якості повітря в місті Мадрид, що включає щоденні дані, була використана для спрощення програмних обчислень у випадках, коли розглядався показник, що відповідає середньому рівню забрудненості повітря відповідним забруднювачем, за день. Це значення було розраховано як усереднене на основі відповідних погодинних даних за день.

Обидві ці вибірки даних містять дані про забрудненість повітря, що вимірювалось за 24 станціями, розташованими в різних локаціях у місті Мадрид. Вимірювання відбувалось за 17 забруднювачами. Не кожна станція оснащена датчиками для вимірювання рівня кожного з 17 забруднювачів. Окрім того не кожен датчик працював протягом всього періоду з 1 січня 2001 року до 30 вересня 2022 року включно, тобто в даних можуть зустрічатися певні пропущені вимірювання. До складу 17 забруднювачів належать:

– діоксид сірки, позначений у подальшому як SO_2;
– моноксид вуглецю (CO);
– оксид азоту (NO);
– діоксид азоту (NO_2);
– забруднення частинками діаметром до 2,5 мм (PM2.5);
– забруднення частинками діаметром до 10 мм (PM10);
– оксиди азоту (NOx);
– озон (O3);
– толуол (TOL);
– бензол (BEN);
– етилбензол (EBE);
– метаксилол (MXY);

– параксилол (PXY);
– ортоксилол (OXY);
– гексан (TCH);
– метан (CH4);
– неметанові вуглеводні (NMHC).

У даній роботі акцент зроблений на прогнозування діоксиду азоту, розглядаючи при цьому інші забруднювачі. Діоксид азоту – це забруднювач, який має значний вплив на стан здоров'я людей, відзначається відносно низьким поточним результатом прогнозування, тобто в розрізі проблеми прогнозування концентрації є достатньо складним, до того ж діоксид азоту пов'язаний з наявним у місті трафіком, а тому це дозволяє продемонструвати логіку, яку було описано в попередньому підрозділі, і яку буде в подальшому дослідженні застосовано саме через зв'язок задач, однією з яких є задача, що стосується трафіку в місті. Окрім того саме в розрізі міста Мадрид діоксид азоту є одним з проблемних забруднювачів, за яким спостерігається перевищення.

Тож загалом вибір саме діоксиду азоту для прогнозування їхнього рівня є достатньо очевидним.

Діоксид азоту в даній вибірці даних представлений вимірами, виконаними за 24 станціями (табл. 4.6). Збір даних за ними розпочався у різний час, тому кількість значень відрізняється. Накопичено ці дані за вибіркою даних стосовно якості повітря в місті Мадрид, що включає щоденні дані.

За кожною позицією у вибірці встановлено відмітки V або N, які відповідно визначають, що дані верифіковані (дійсні) або ні. Якщо дані не верифіковані, то до результуючого набору було внесено порожнє значення (NaN). За кожною станцією було проаналізовано відсоток таких порожніх (пропущених значень). Розрахунок цього відсотку виконувався через обчислення відсоткового співвідношення між кількістю пропущених значень разом з кількістю позицій (дат), які взагалі не були вказані для станції (наприклад, якщо протягом якогось періоду вимірювання через ті чи інші причини не виконувались), та кількістю днів, які пройшли від початку вимірювань за цією станцією до закінчення.

Таблиця 4.6 – Станції, за якими відбувалося вимірювання рівня діоксиду азоту в повітрі

Назва станції	Початок збору даних	Всього записів	Порожні записи, шт.	Відсоток пропущених записів, %	Найменше значення, мкг/м^3	Найбільше значення, мкг/м^3
1	2	3	4	5	6	7
Plaza de Espana	2001-01-01	7823	193	3,94	2	175
Escuelas Aguirre	2001-01-01	7486	144	7,57	7	172
Avenida de Ramon y Cajal	2001-01-01	7943	34	0,43	3	159
Arturo Soria	2001-01-01	7881	70	1,66	1	153
Villaverde Alto	2001-01-01	6451	100	20,04	4	151
Calle Farolillo	2001-01-01	7943	142	1,79	3	185
Casa de Campo	2001-01-01	7943	133	1,67	1	119
Plaza del Carmen	2001-01-01	7943	128	1,61	3	157
Moratalaz	2001-01-01	7943	71	0,89	4	151
Cuatro Caminos	2001-01-01	7943	60	0,76	3	166
Barrio del Pilar	2001-01-01	7853	88	2,24	3	165
Vallecas	2001-01-01	7943	56	0,71	4	135
Barajas	2003-01-01	7213	31	0,43	2	143
Mendez Alvaro	2009-12-01	4687	48	1,02	3	121

Продовження таблиці 4.6

1	2	3	4	5	6	7
Ensanche de Vallecas	2009-12-01	4687	20	0,43	4	121
Sanchinarro	2009-11-01	4717	49	1,04	2	131
El Pardo	2009-12-01	4687	13	0,28	1	74
Parque Juan Carlos I	2009-12-01	4687	52	1,11	1	87
Paseo de la Castellana	2010-06-01	4505	16	0,36	2	134
Parque del Retiro	2010-01-01	4656	19	0,41	2	111
Plaza de Castilla	2010-02-01	4625	71	1,54	2	149
Urbanizacion Embajada (Barajas)	2010-01-01	4656	41	0,88	1	119
Plaza Elíptica	2010-01-01	4656	62	1,33	4	145
Tres Olivos	2010-01-01	4656	53	1,14	2	131

Як видно з таблиці 4.6, достатньо велика кількість значень (більше 20 %) пропущена за станцією Villaverde Alto, однак у випадку дослідження, починаючи з 2018 року, цей відсоток є значно меншим – лише 0,29 % (або 0,37 %, починаючи з 2019 року).

Звичайно, що таблиця 4.6 лише узагальнює дані вибірки, але важливими є і окремі значення, тому далі було розглянуто графіки, які демонструють зміну значень протягом всього періоду спостережень для цих станцій.

На рис. 4.30-4.35 представлено за кожною зі станцій значення рівня забрудненості атмосферного повітря діоксидом азоту на основі розрахованих середньомісячних значень, тобто кожна точка – це середнє значення серед всіх днів місяця, виключаючи пропущені значення.

Представлена логіка була використана і посилена за рахунок того, що інтерполяція для пропущених середньомісячних значень не виконувалась. Тобто фактично за цими графіками можна побачити пропущені значення, якщо протягом цілого місяця або декількох місяців спостереження не виконувались. Наприклад, для вищеописаної станції Villaverde Alto на рис. 4.31 на графіку а) помітно період, протягом якого спостереження не виконувались.

За графіками на рис. 4.30-4.35 також помітно, що загальна форма функції може значно відрізнятися від станції до станції, що також вірно і для амплітудних значень, які можуть значно відрізнятися від станції до станції, таким чином фактично характеризуючи певні зони, де викиди діоксиду азоту є більш значними. Загалом все це вказує на необхідність окремо досліджувати рівень забрудненості діоксидом азоту для різних станцій, щоб перевірити запропоновані фактично моделі. Робити висновок за окремою станцією або навіть декількома, які виділені з загального набору, як це виконується в певних дослідженнях, може бути некоректно і призводити до упередженості.

Якщо користуватися рекомендаціями Всесвітньої організації здоров'я [24], то рекомендованим середньорічним рівнем для діоксиду азоту є 10 мкг/м^3 протягом року, а поетапне зниження передбачається з рівня у 40 мкг/м^3.

Аналіз даних за різними станціями (у вимірі місяців) вказує на те, що з забрудненням повітря діоксидом азоту в Мадриді спостерігаються проблеми, хоча і за всіма станціями спостерігалось фактично певне зменшення граничних рівнів коливання середнього значення в останні роки.

Окрім того Всесвітньою організацією здоров'я визначено також і критичні рівні для короткострокового впливу протягом 24 годин: рекомендоване значення – 25 мкг/м^3, а зниження розпочинається зі 120 мкг/м^3. Фактично ці рекомендації є основою прийняття рішень при медичному діагностуванні. Але в той же час мають враховуватися індивідуально в залежності від стану здоров'я людини.

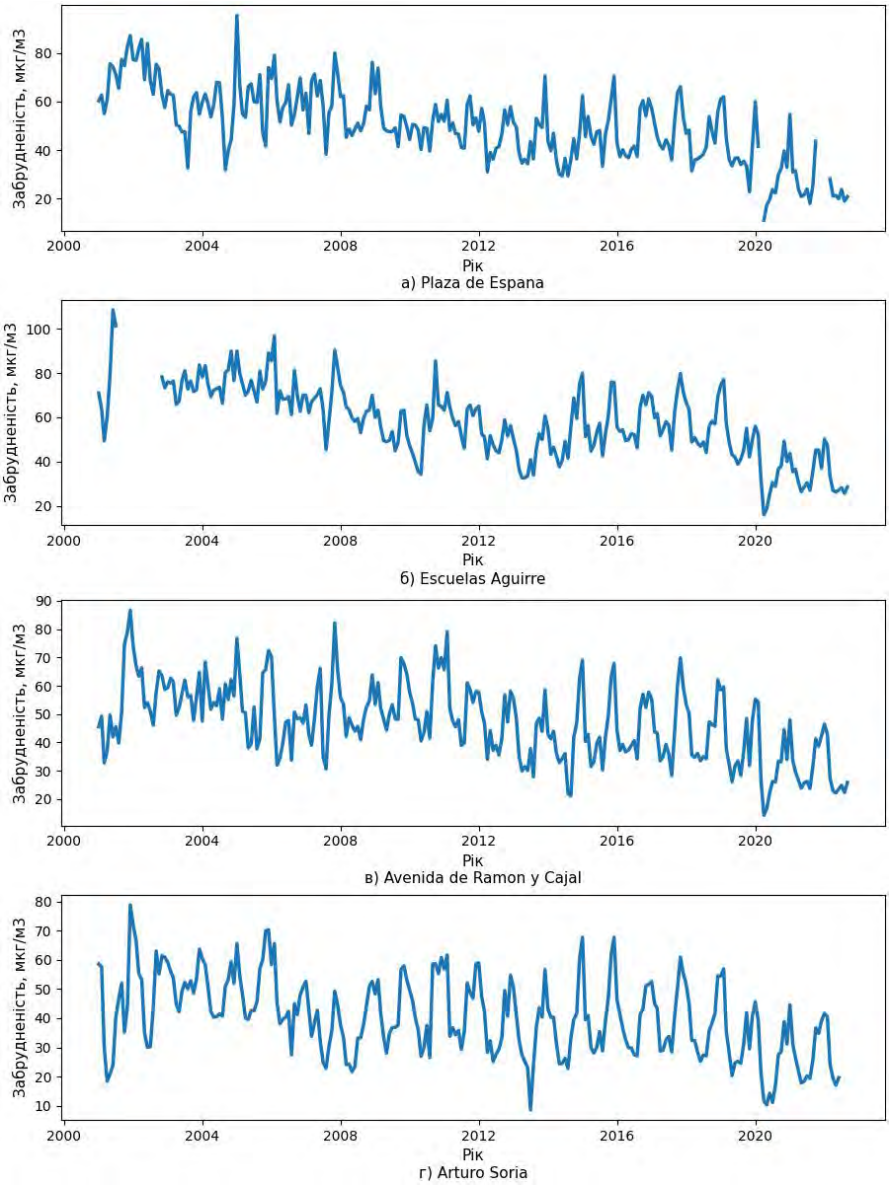


Рисунок 4.30 – Рівень забрудненості атмосферного повітря діоксидом азоту за представленими середньомісячними значеннями

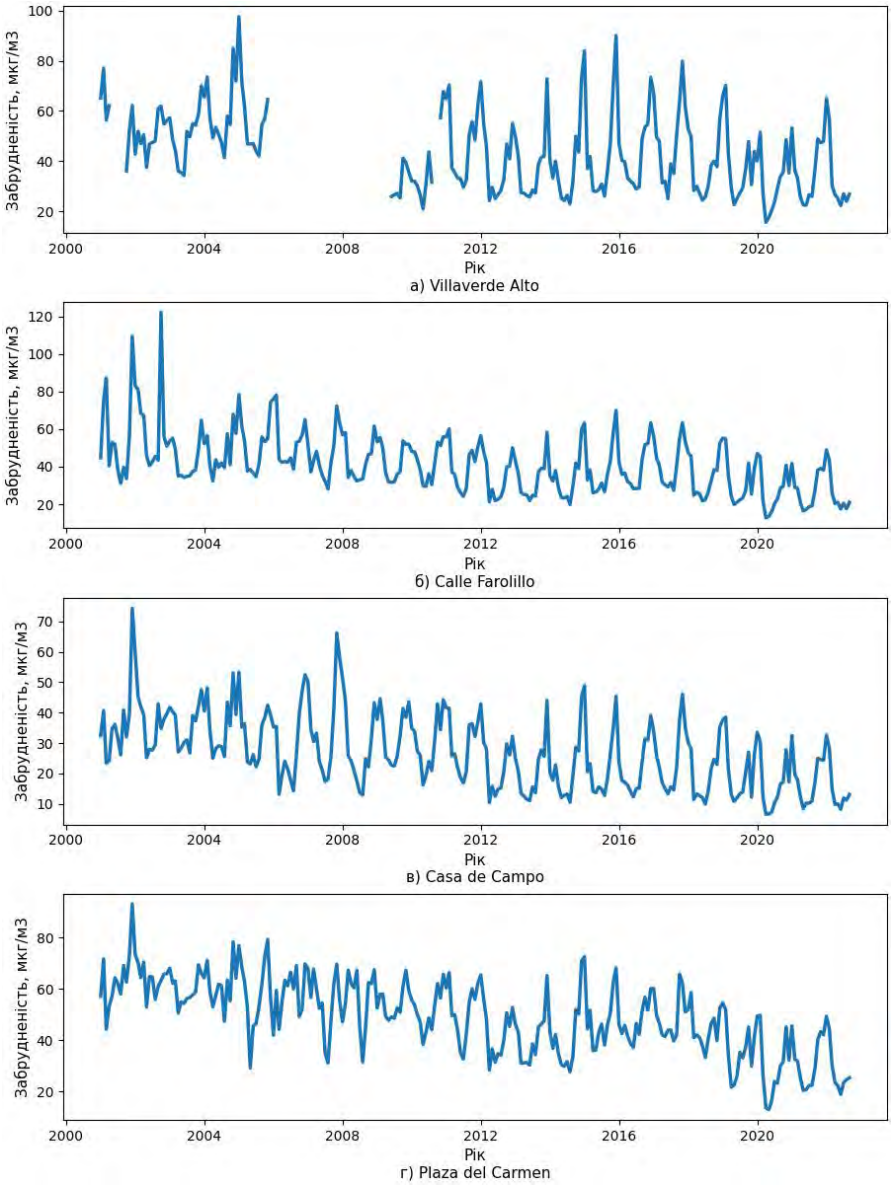


Рисунок 4.31 – Рівень забрудненості атмосферного повітря діоксидом азоту за представленими середньомісячними значеннями

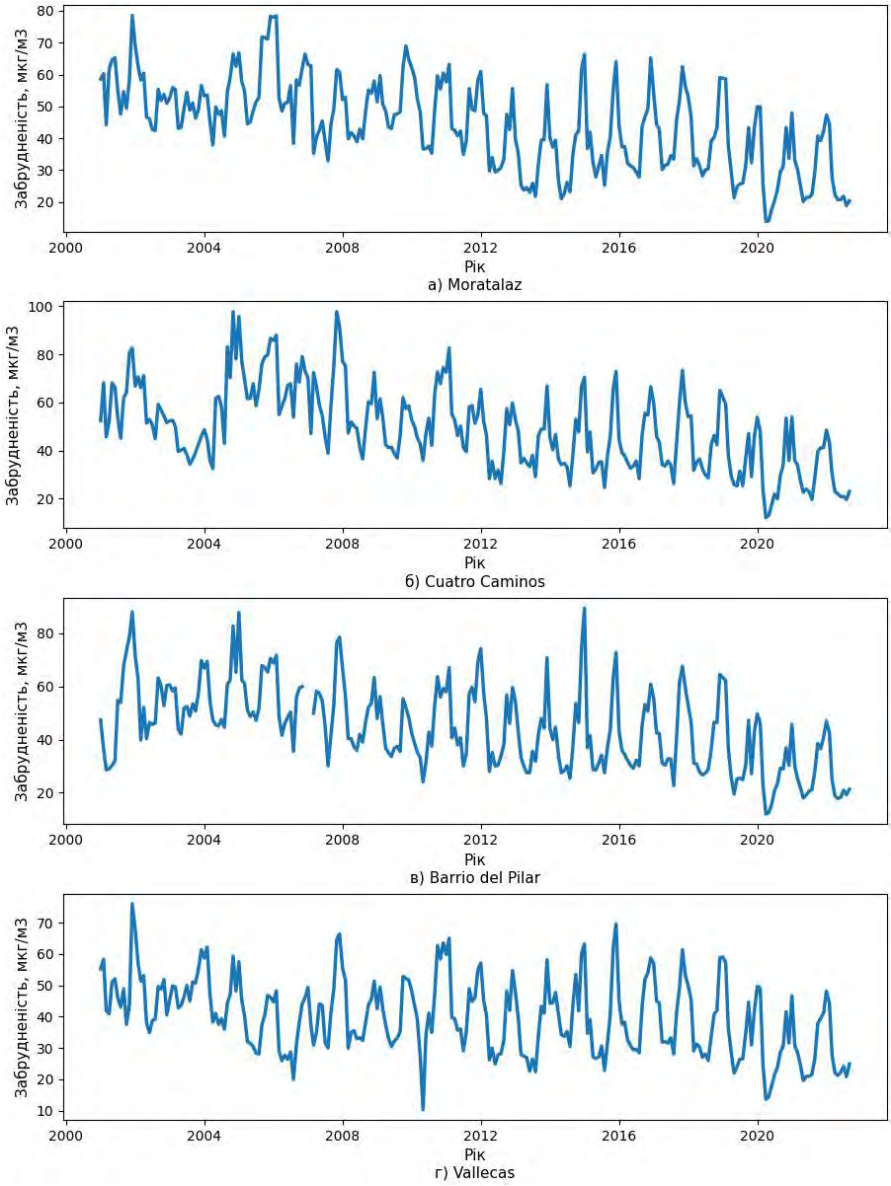


Рисунок 4.32 – Рівень забрудненості атмосферного повітря діоксидом азоту за представленими середньомісячними значеннями

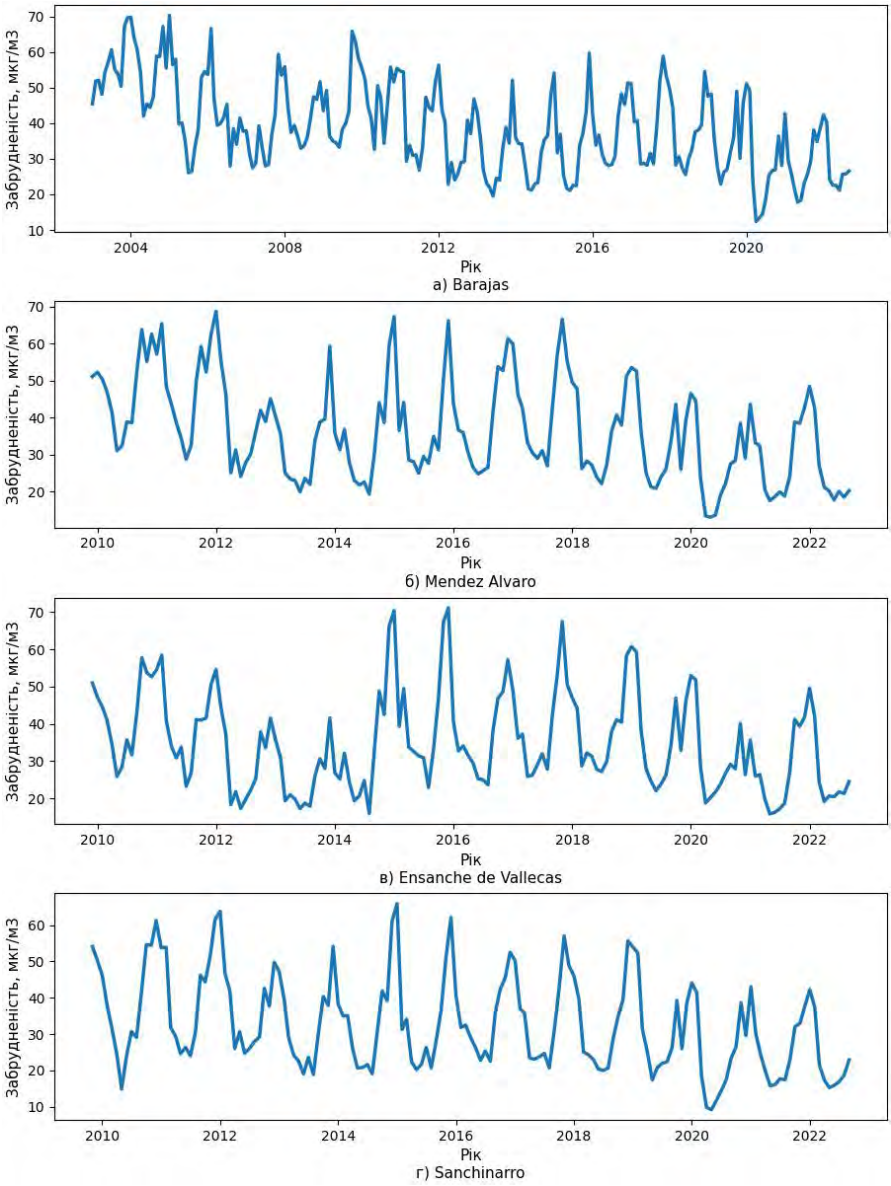


Рисунок 4.33 – Рівень забрудненості атмосферного повітря діоксидом азоту за представленими середньомісячними значеннями

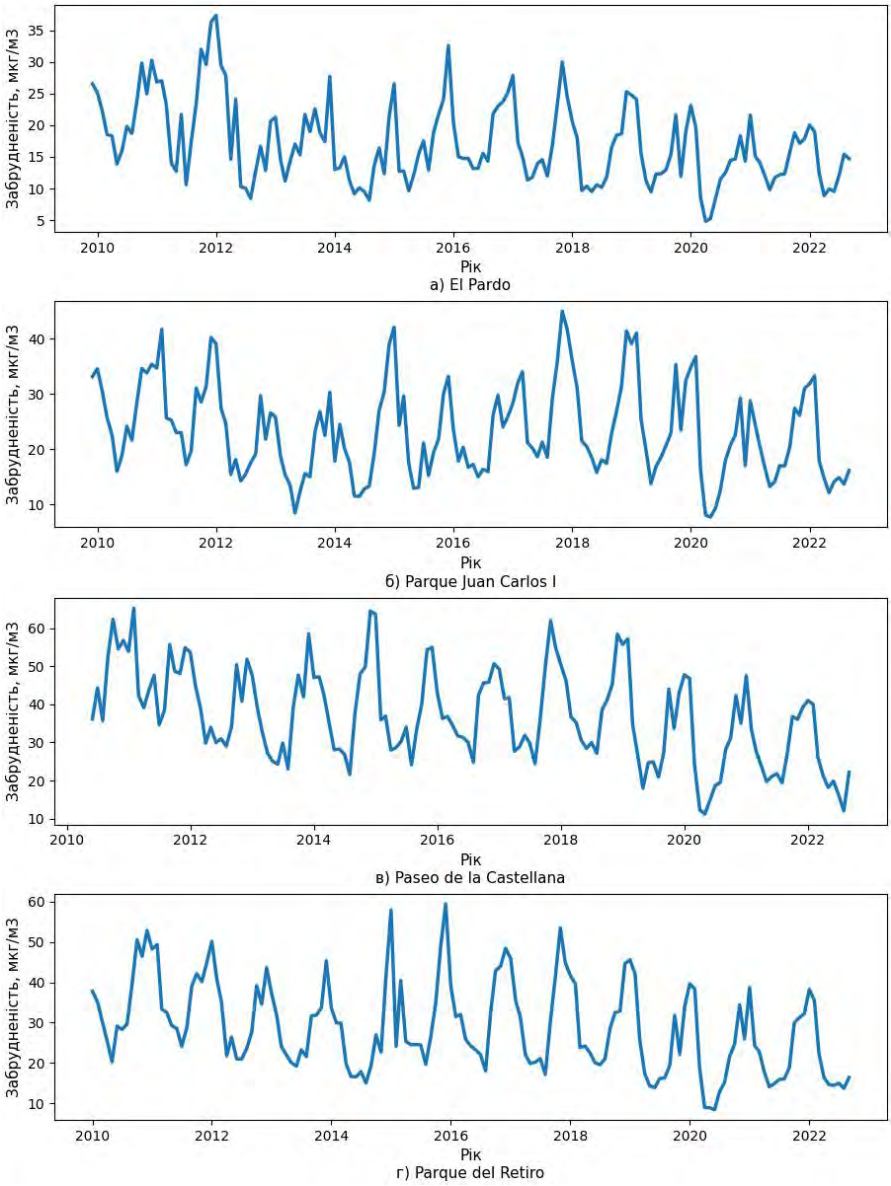


Рисунок 4.34 – Рівень забрудненості атмосферного повітря діоксидом азоту за представленими середньомісячними значеннями

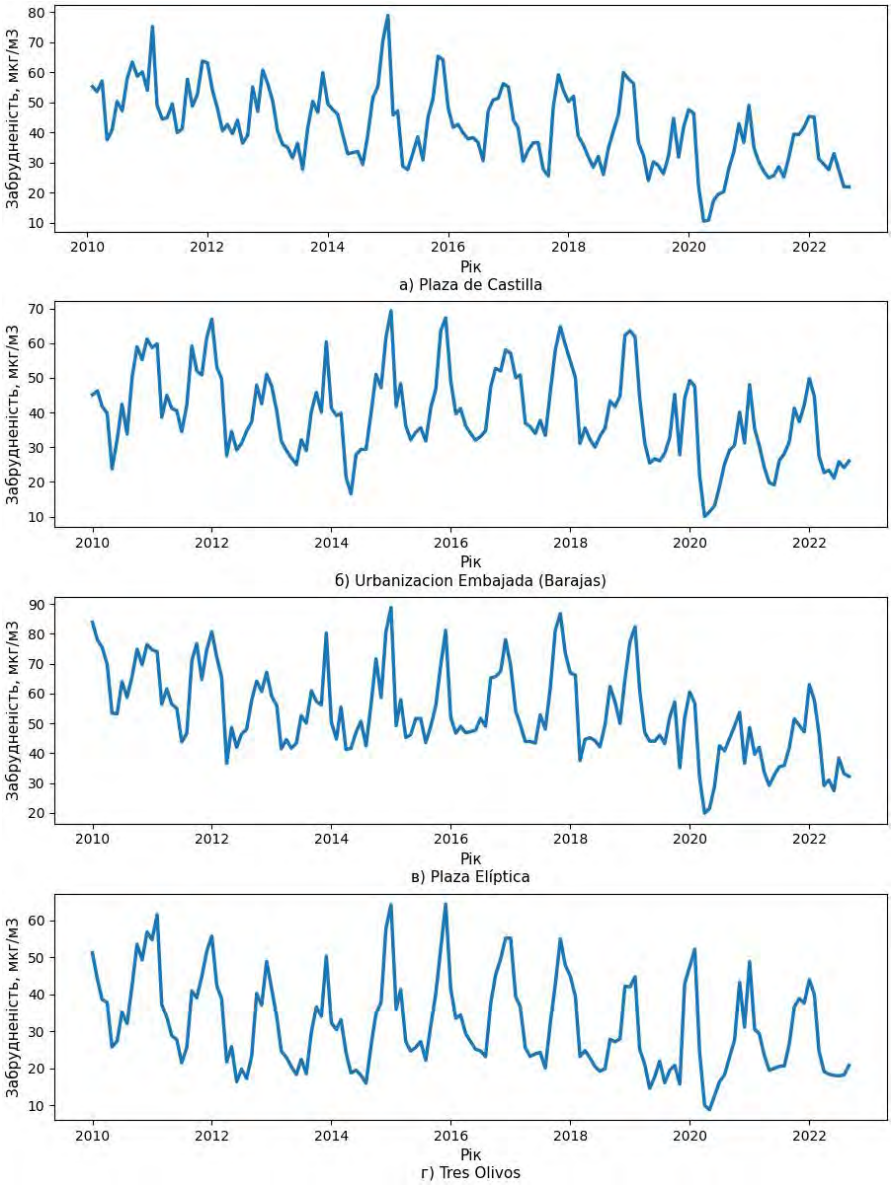


Рисунок 4.35 – Рівень забрудненості атмосферного повітря діоксидом азоту за представленими середньомісячними значеннями

Для більш детального вивчення співвідношень між рівнем забрудненості атмосферного повітря діоксидом азоту було представлено місячні значення в розрізі року для одразу декількох станцій, щоб продемонструвати співвідношення не тільки між місяцями, але і між різними локаціями (рис. 4.36). Слід відзначити, що відсутність значень для станції Plaza de Espana для деяких місяців була перевірена зокрема і за щогодинними даними. Було виявлено, що тільки протягом березня 2020 року за 2 дні було зафіксовано деякі значення у окремі години, що звичайно не дозволяє встановити середньомісячний показник, тому позначка про недостовірність цих даних у щоденних даних є абсолютно коректною.

Графіки на рис. 4.36 демонструють, що рівень забрудненості може значно відрізнятися на різних станціях. При цьому можна явно спостерігати, що значення в місяцях на початку року та в кінці року є вищими ніж всередині року. Однак форма функціональної залежності, яка демонструє цю зміну для даних 3 станцій значно відрізняється. Також на графіку для 2020 року видно, що з початком пандемії COVID-19 рівень забрудненості повітря діоксидом азоту значно зменшився, продовживши демонструвати нижчі рівні протягом наступних місяців, порівняно з 2019 роком, наприклад. Хоча поступово далі рівень вмісту діоксиду азоту в повітрі збільшується. Проте вже в листопаді цей ефект точно знятий, адже рівень забрудненості повітря діоксидом азоту за всіма 3 станціями в цей місяць перевищує рівень попереднього року. Тож на графіку можна побачити, що зниження ділової активності протягом пандемії COVID-19 вплинуло на екологічну ситуацію в Мадриді щонайменше в розрізі забрудненості повітря діоксидом азоту.

Проте вказане вище і загалом коректне твердження про те, що найчастіше значення забрудненості на початку року та в кінці є дещо вищими, виконується не завжди. Трапляються випадки, коли наприклад, на початку року таке домінування є достатньо невираженим, хоча може значно відрізнятися для різних станцій. Такий випадок представлено на рис. 4.37. У 2016 році для станції Plaza Eliptica перевага початкових місяців року за рівнем забрудненості повітря діоксидом азоту фактично відсутня, що характерно і для багатьох інших станцій, а для станції Arturo Soria зниження цього показника всередині року помітно як порівняно з

початком, так і з кінцем року. Приклади таких залежностей у інші роки приведені на рис. 4.38-4.40.

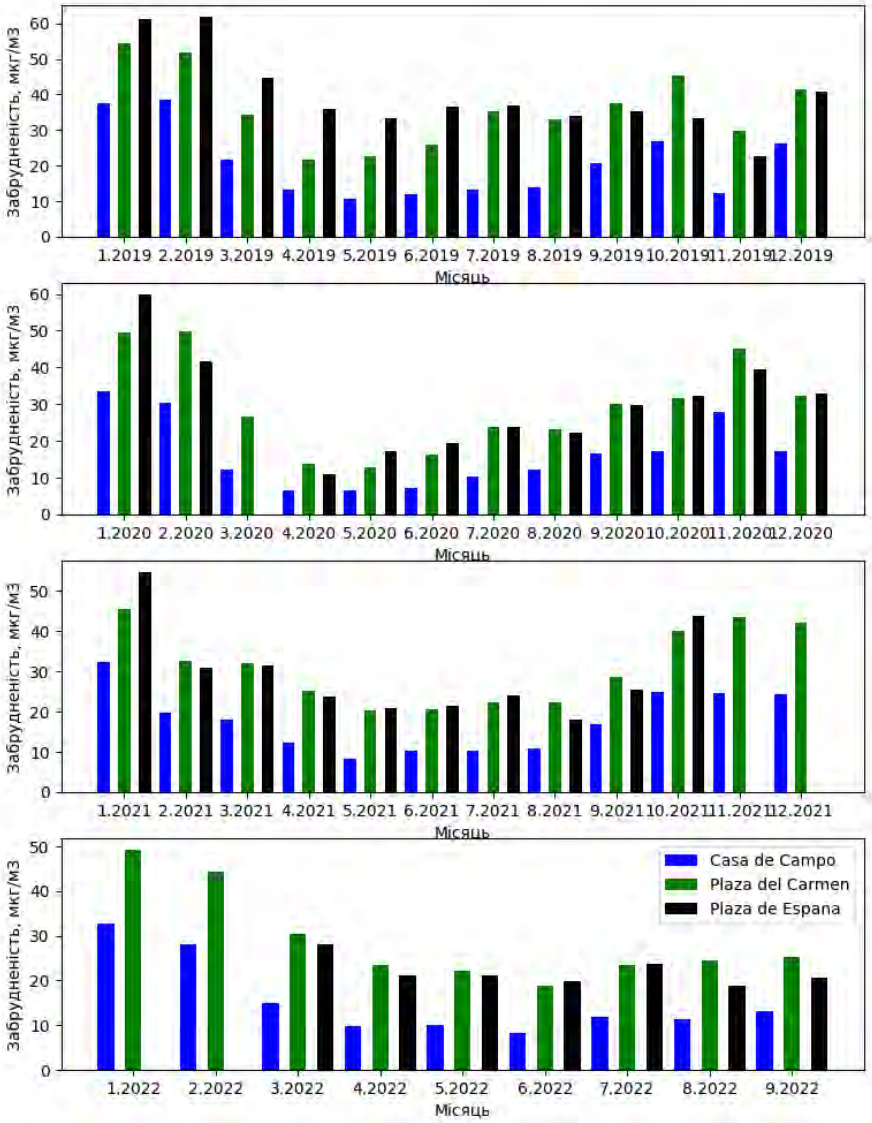


Рисунок 4.36 – Середньомісячний рівень забрудненості атмосферного повітря діоксидом азоту протягом 2019-2022 років

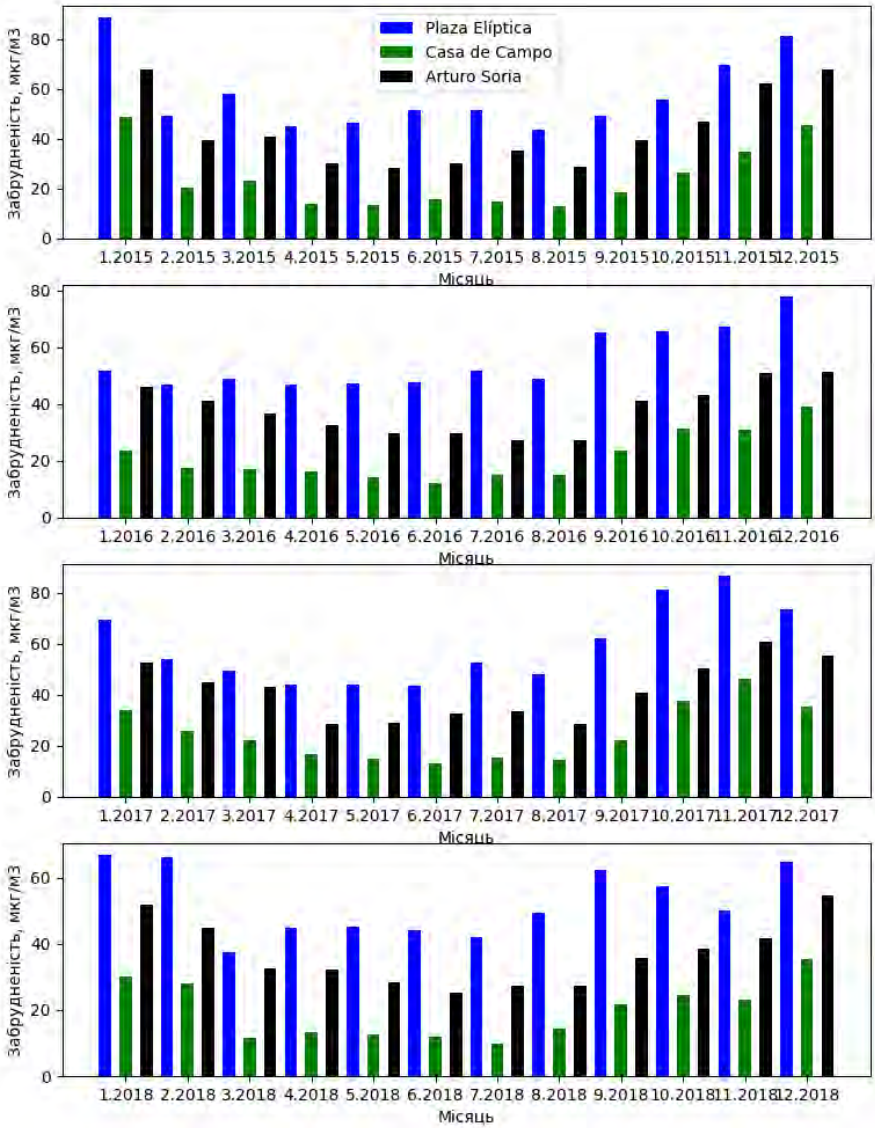


Рисунок 4.37 – Середньомісячний рівень забрудненості атмосферного повітря діоксидом азоту протягом 2015-2018 років

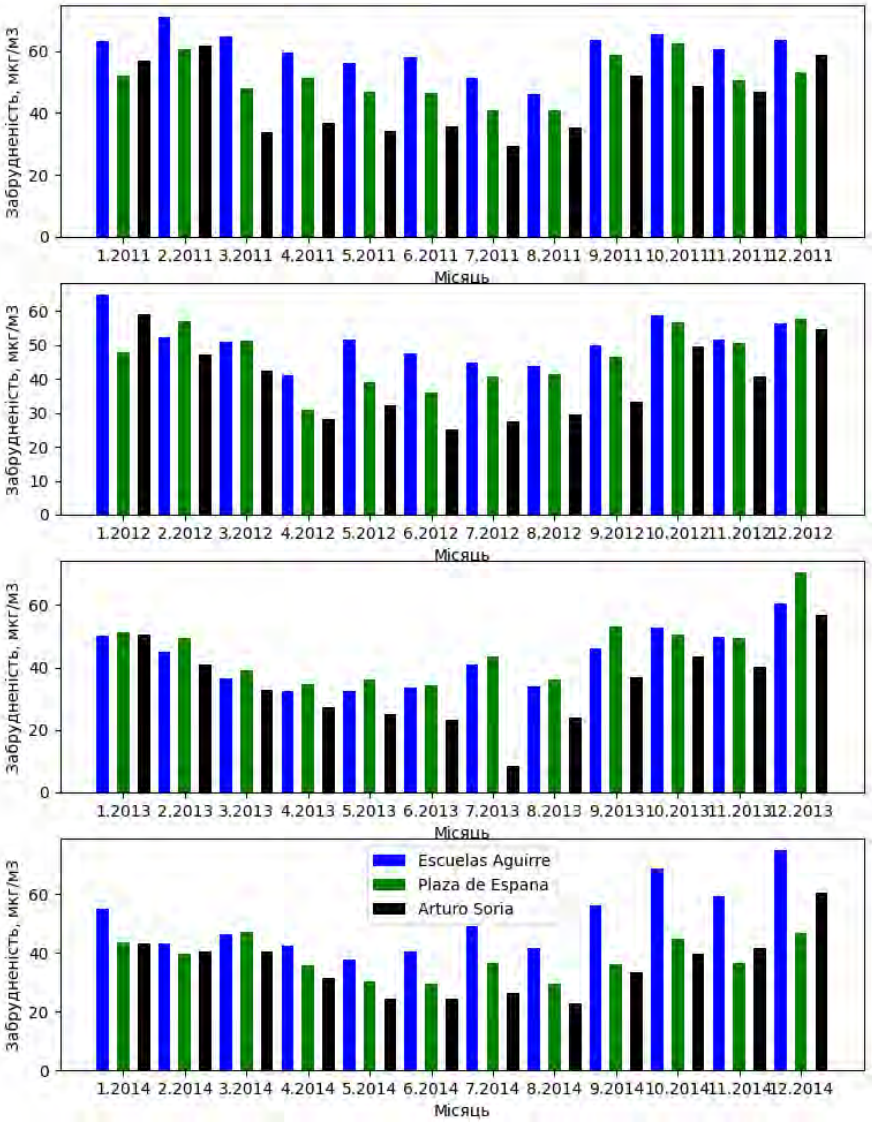


Рисунок 4.38 – Середньомісячний рівень забрудненості атмосферного повітря діоксидом азоту протягом 2011-2014 років

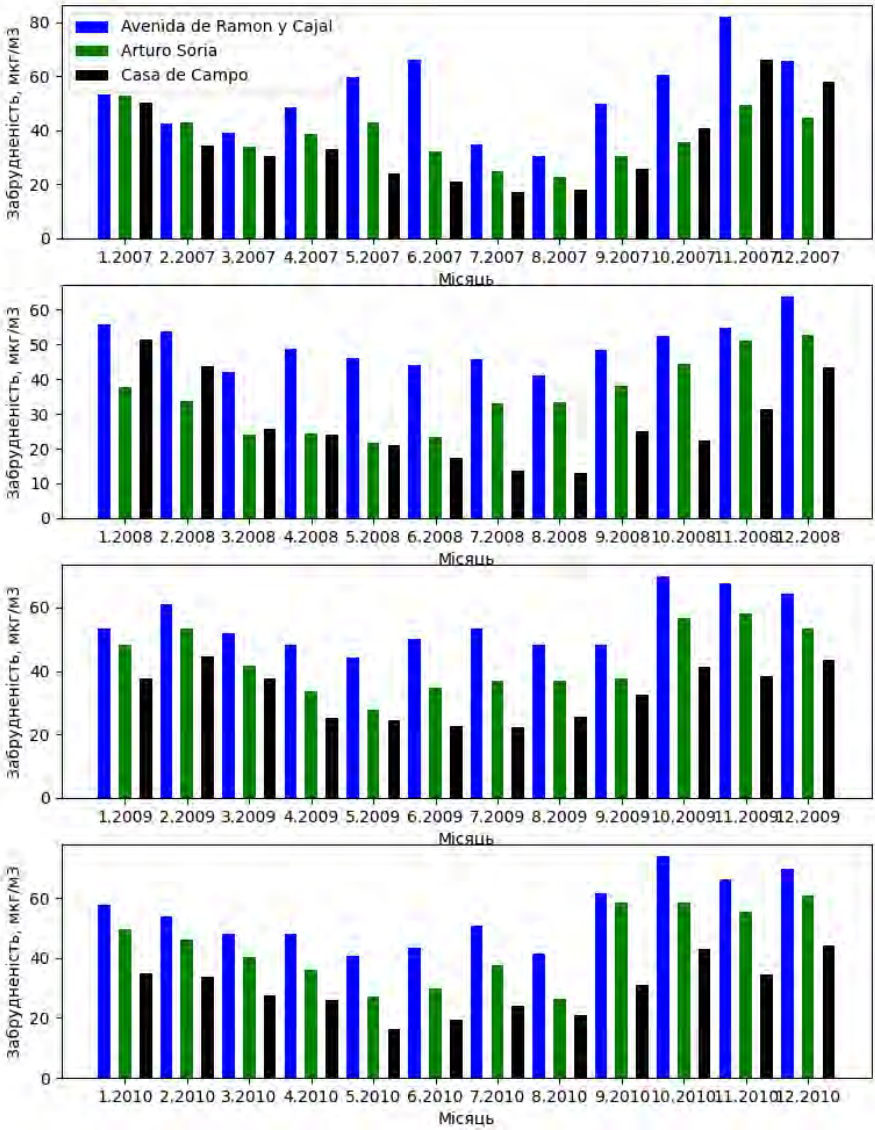


Рисунок 4.39 – Середньомісячний рівень забрудненості атмосферного повітря діоксидом азоту протягом 2007-2010 років

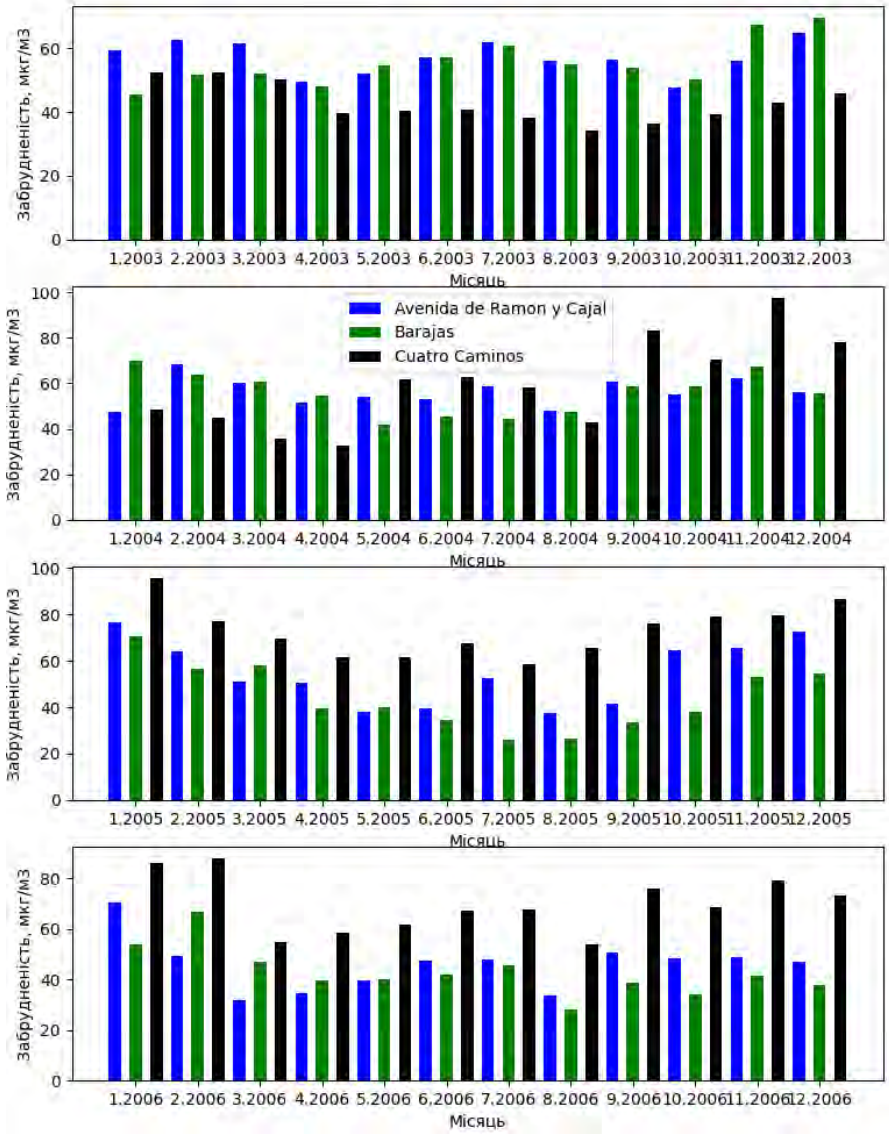


Рисунок 4.40 – Середньомісячний рівень забрудненості атмосферного повітря діоксидом азоту протягом 2003-2006 років

Представлені графіки також вказують на те, що швидше за все на рівень забрудненості повітря діоксидом азоту впливають різні фактори і пояснити отримані значення, наприклад, сезонністю і деякими додатковими постійними для різних станцій факторами не можна. Цей вплив важливо дослідити.

Розглянуті вище рекомендаційні рівні вмісту діоксиду азоту адекватно розглядати звичайно за відповідними часовими проміжками, тобто спираючись на річний рівень та на погодинні рівні, щоб виділити проміжок у 24 години. Тому спочатку розглянемо, який середньорічний рівень забрудненості атмосферного повітря діоксидом азоту був характерний за весь період спостереження для кожної зі станцій.

На рис. 4.41-4.43 продемонстровано середньорічний рівень забрудненості атмосферного повітря діоксидом азоту на всіх розглянутих вище станціях. Ці значення обчислено шляхом усереднення всіх значень, які спостерігались протягом днів року, тобто порожні значення були перед тим відкинуті. На графіках окремо виділені для зручності візуалізації описані вище рівні у 40 мкг/м^3 (червоним кольором) та 10 мкг/м^3 (жовтим кольором).

Як видно, протягом 3 останніх років, тобто 2020-2022 (2022 рік представлений тільки 9 місяцями), концентрація діоксиду азоту в повітрі була доведена до рівня, нижче червоного: хоча за станцією Plaza Elíptica, що є одним з активних транспортних хабів, у 2020 році ще спостерігалось значення у 40,56 мкг/м^3, у 40,6 мкг/м^3 у 2021 році та трохи нижче, 39,78 мкг/м^3, за неповний 2022 рік.

З іншого боку можна побачити, що за період до 2017 року включно за багатьма станціями за результатами багатьох років спостерігались значення, що перевищували критичний рівень. Тож загалом можна вказати на те, що незважаючи на те, що Мадрид є одним з економічно розвинених міст з сучасним підходом до впорядкування простору, певні екологічні проблеми характерні для міста, а тому визначення рекомендацій для поведінки різних категорій людей відносно їхнього стану здоров’я є важливим. Однією з таких екологічних проблем є концентрація діоксиду азоту.

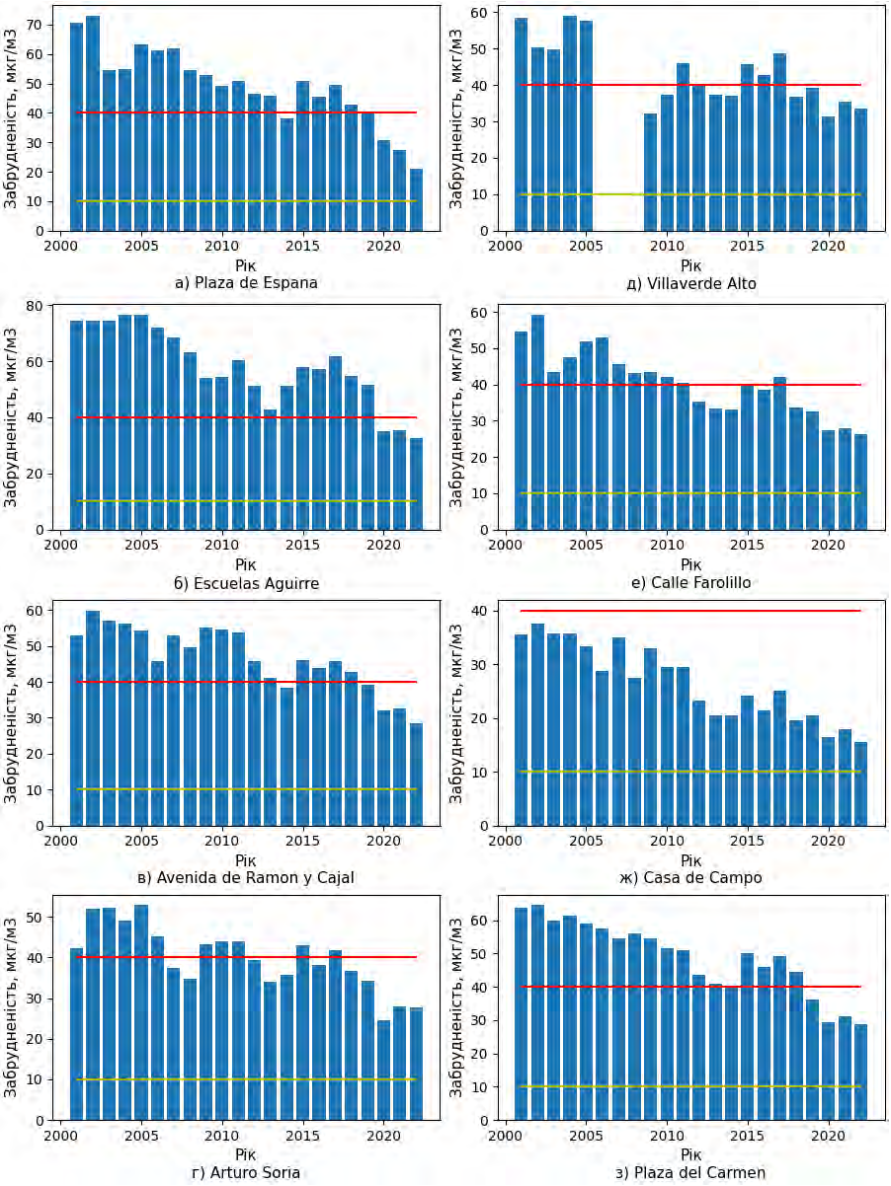


Рисунок 4.41 – Середньорічний рівень забрудненості атмосферного повітря діоксидом азоту

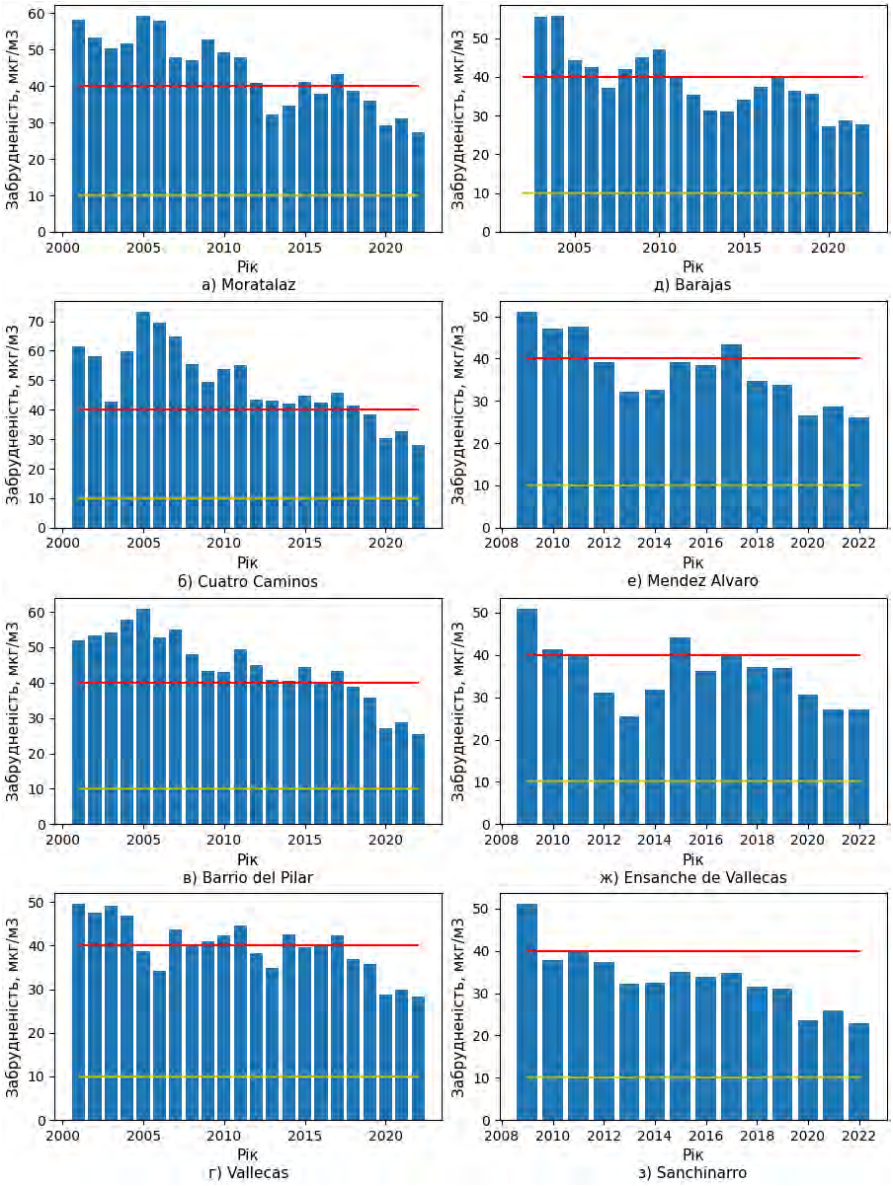


Рисунок 4.42 – Середньорічний рівень забрудненості атмосферного повітря діоксидом азоту

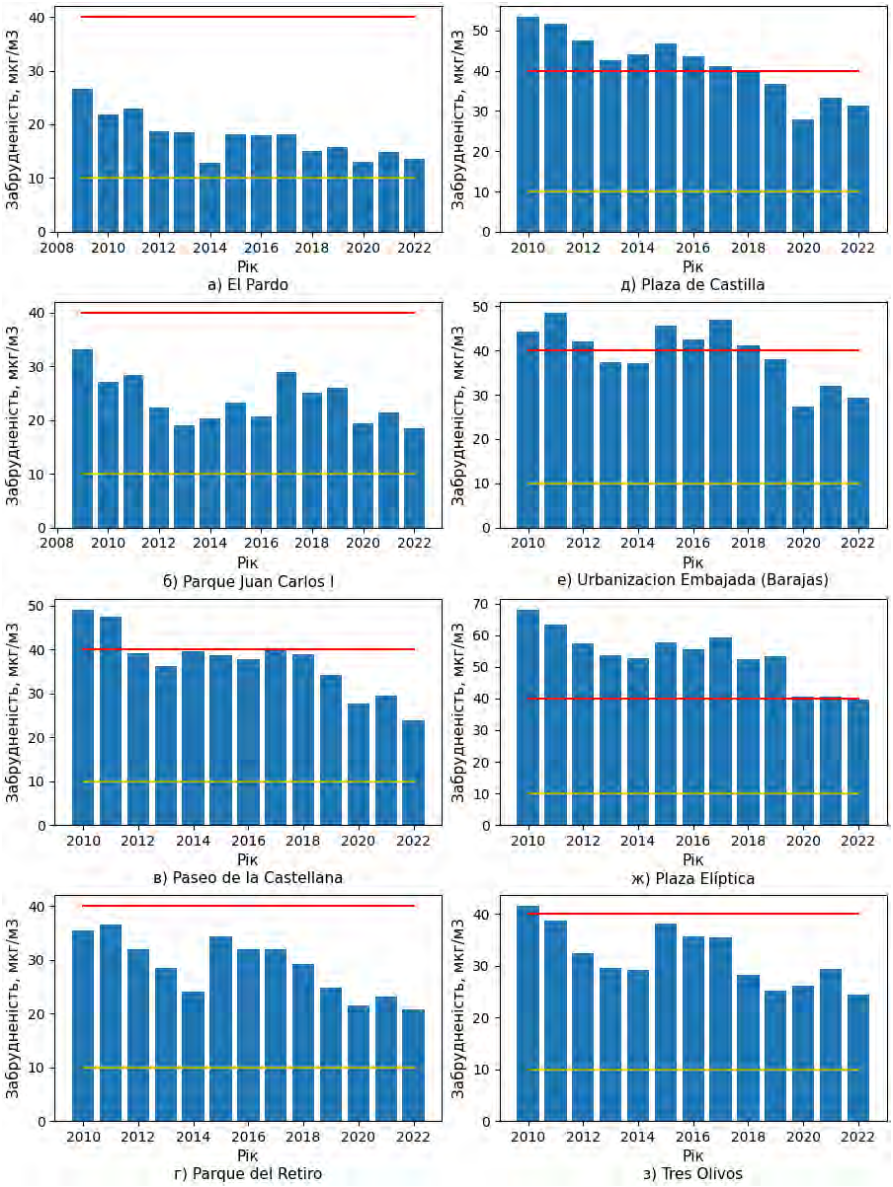


Рисунок 4.43 – Середньорічний рівень забрудненості атмосферного повітря діоксидом азоту

Якщо вивчати більш детально зміни найменшого за тривалістю рівня забрудненості повітря діоксидом азоту, то потрібно використовувати вибірку з щоденними даними. Приклади таких залежностей для різних станцій протягом місяця з використанням значень, що відповідають кожній годині, приведені на рис. 4.44-4.45. Для цих графіків використано ті значення, які були наявні у вибірці. Тому якщо певні значення пропущені, то такі точки на графіку не виводились. Оскільки в місячному розрізі використання інтерполяції на графіках дещо спотворювало певні значення або викривляло функціональну залежність, то був використаний саме такий підхід. Разом з даними з таблиці 4.6 це додатково демонструє необхідність коректної обробки даних перед їх використанням у моделях.

Приклади залежностей на рис. 4.44-4.45 демонструють, що для деяких станцій у один і той же період значення можуть достатньо суттєво відрізнятися (рис. 4.44), у той час як для інших вони можуть знаходитися часто дуже близько, що на прикладах на рис. 4.45 призводило до певного накладання ліній. При цьому з цих графіків помітно, що відбуваються певні коливання за кожного спостереження, тобто щогодинно. Проте ці графіки дещо ускладнюють сприйняття даних, адже важко помітити чи ці коливання відбуваються в один і той же самий момент в одних і тих самих напрямках. Це детальніше можна розглянути на прикладах графіків, на яких винесено дані за декілька днів (в основному за 3 дні). Приклади таких графіків приведено на рис. 4.46-4.51.

Графіки на рис. 4.46-4.51, зважаючи на їхні характеристики, побудовано за дещо відмінним принципом від попередніх. На них виконано кубічну інтерполяцію, що дозволило з одного боку представити більш плавні відображення, а з іншого боку замінити пропущені значення, якщо вони траплялись.

Графік на рис. 4.46 чітко демонструє, що значення, форма самої функціональної залежності та періоди змін напрямків можуть достатньо суттєво відрізнятися для різних станцій. На цих графіках під діленнями підписано номер місяця з номером дня місяця через дефіс та з додаванням через пробіл кількості годин, що відповідає цій відмітці.

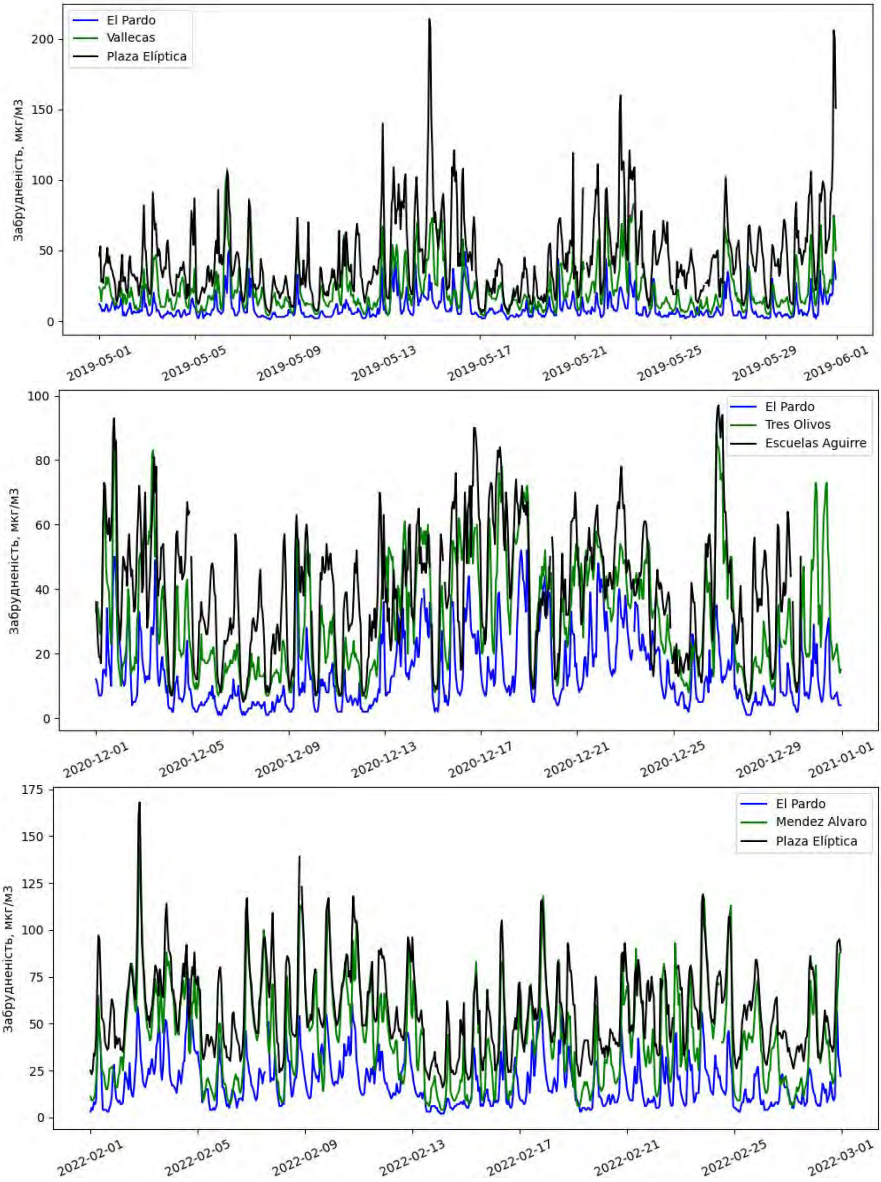


Рисунок 4.44 – Рівень забрудненості атмосферного повітря діоксидом азоту протягом місяця

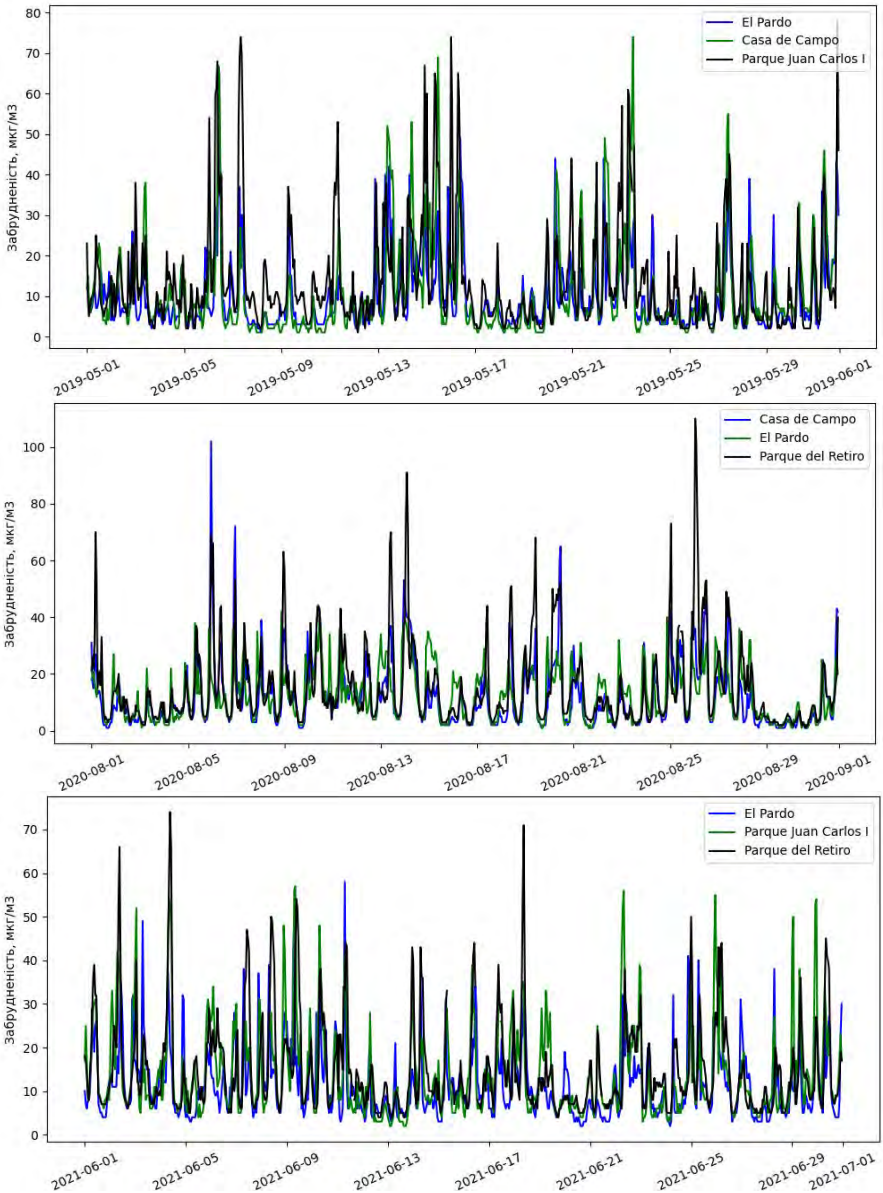


Рисунок 4.45 – Рівень забрудненості атмосферного повітря діоксидом азоту протягом місяця

Графік на рис. 4.46 демонструє, що в той час, як значення забрудненості може підійматися на одній станції, воно може зменшутивася на іншій. У той же час моменти зміни цих напрямків (збільшення або зменшення) можуть не співпадати для різних станцій (пам'ятаючи при цьому, що на графіках виведені погодинні значення, тому одному діленню відповідає 6 годин, тобто тільки 6 окремих значень).

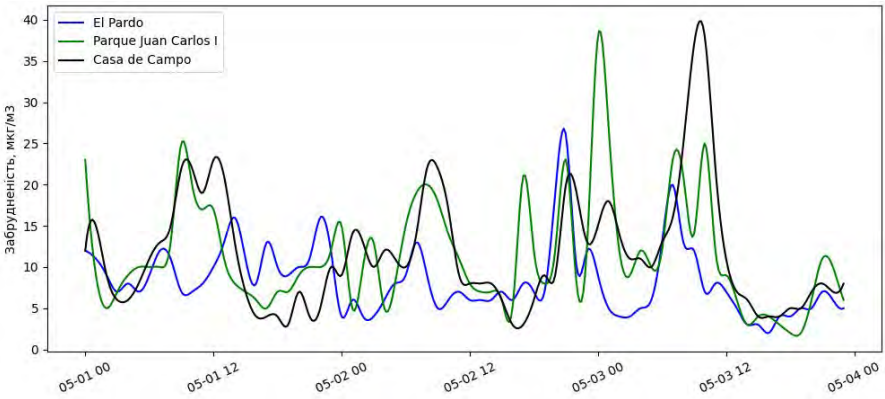


Рисунок 4.46 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 1-3 травня 2019 року

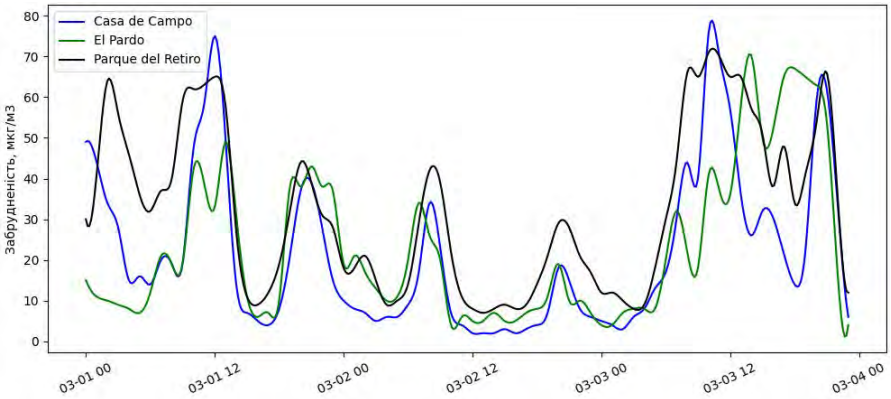


Рисунок 4.47 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 1-3 березня 2022 року

Рис. 4.47 у свою чергу демонструє графік, на якому протягом щонайменше 1,5 діб можна спостерігати дуже близькі значення забрудненості повітря діоксидом азоту одразу на 3 станціях. Хоча вони звичайно не повністю співпадають, а моменти зміни напрямків також не повністю співпадають.

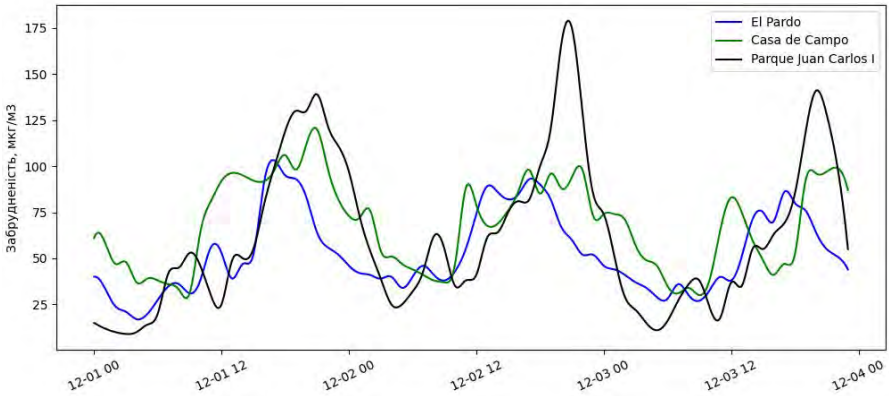


Рисунок 4.48 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 1-3 грудня 2015 року

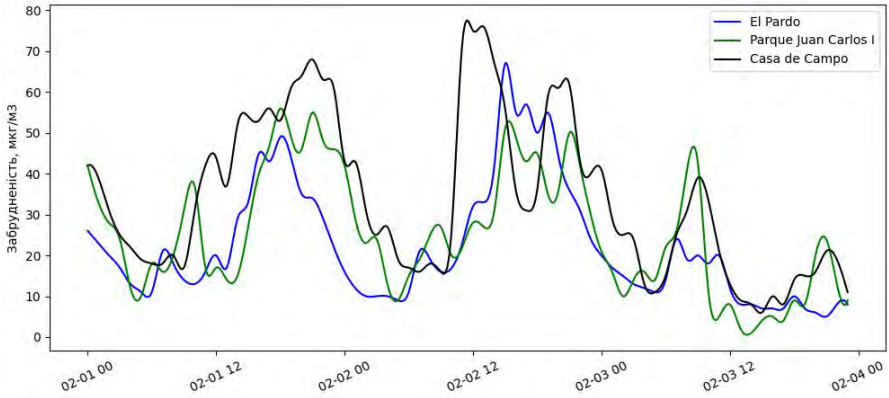


Рисунок 4.49 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 1-3 лютого 2016 року

Вище було вказано також на важливість короткочасного впливу діоксиду азоту. Короткочасний вплив визначається проміжком у 24 години. З цього боку цікавим спостереженням є період у січні 2021

року, коли як можна чітко побачити на рис. 4.50, рівень забрудненості повітря діоксидом азоту протягом 16 січня на станції Avenida de Ramon y Cajal не знижувався нижче 100 мкг/м^3. Це не критичний рівень, хоча в певні періоди вміст діоксиду азоту в повітрі значно перевищує і критичний рівень (рис. 4.51), але це вже рівень вище порогового, тобто він фактично призводить до підвищення смертності.

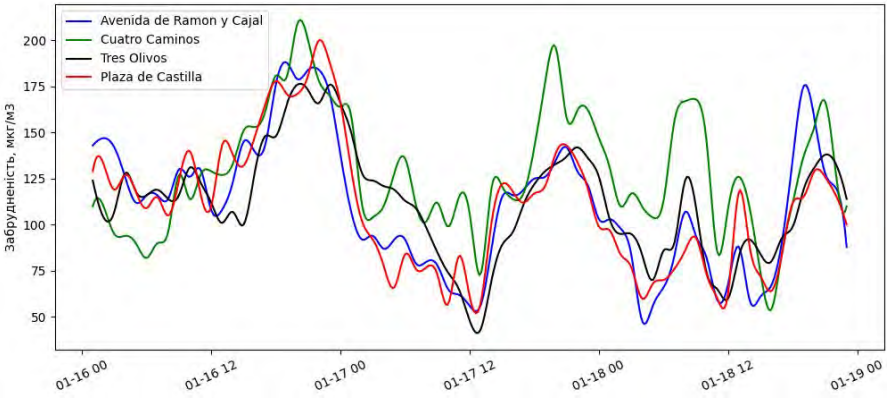


Рисунок 4.50 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 16-18 січня 2021 року

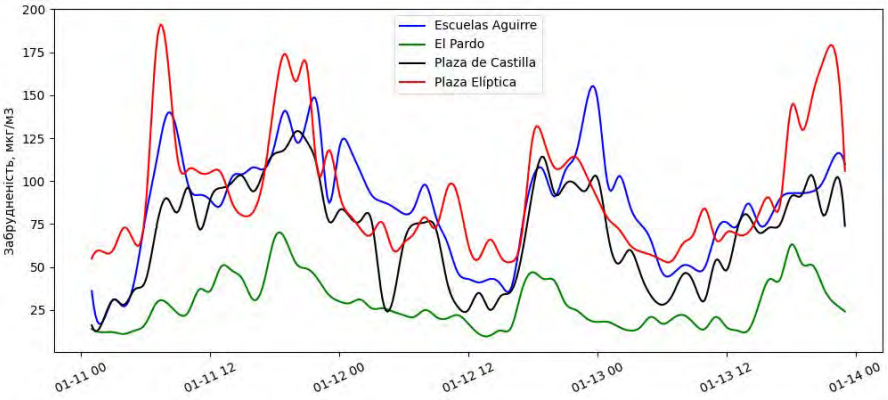


Рисунок 4.51 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 11-13 січня 2019 року

З розглянутих представлень вже можна зробити певні висновки про

вплив сезонності на забрудненість повітря діоксидом азоту, але для встановлення повної картини доцільно також зконцентруватися на щоденних результатах, а для цього варто розглянути, як змінюється забрудненість повітря в різні дні тижня. Для візуалізації цієї залежності було виділено окремі станції, для кожної з яких представлено декілька прикладів зміни забрудненості повітря в різні години різних днів тижня. Ці приклади приведені на рис.4.52-4.54 для станції Escuelas Aguirre та на рис. 4.55-4.58 для станції Vallecas. При цьому кожна з кривих, представлених на графіках, відображає 7 днів тижня, тобто понеділок, дата якого вказана найвище у легенді графіка, вівторок, середа, четвер, п'ятниця, субота, неділя (вказана останньою). На рис. 4.52, 4.54 помітна тенденція щодо того, що у вихідні дні (суботу та неділю) рівень забрудненості знижується порівняно з робочими днями. Це особливо характерно для денних годин. Проте це не абсолютно завжди виконується, як наприклад, період, приведений на рис. 4.53.

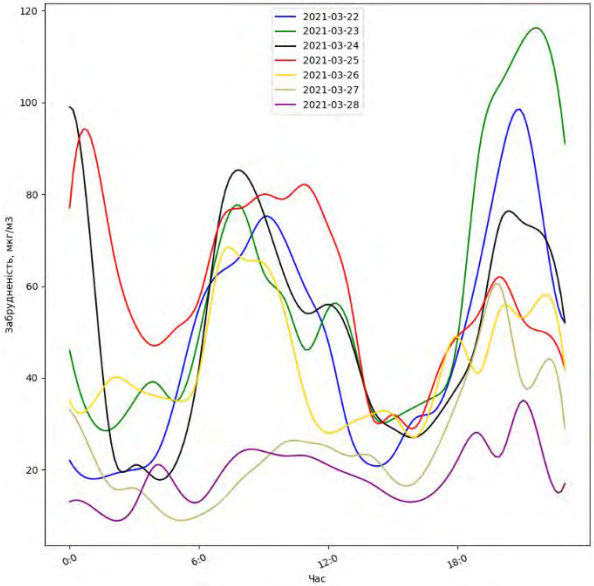


Рисунок 4.52 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 22-28.03.2021 на станції Escuelas Aguirre

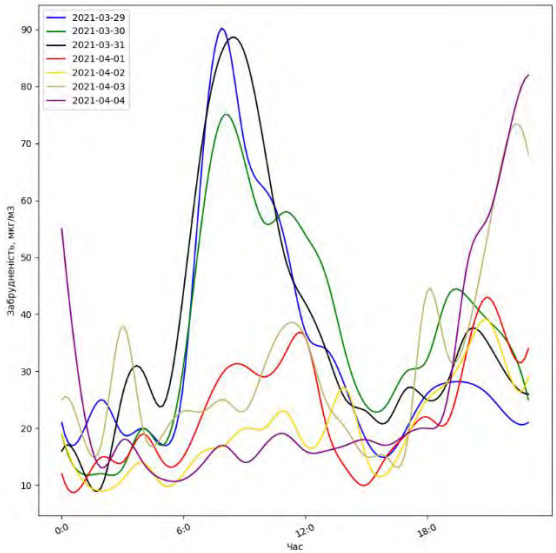


Рисунок 4.53 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 29.03-4.04.2021 на станції Escuelas Aguirre

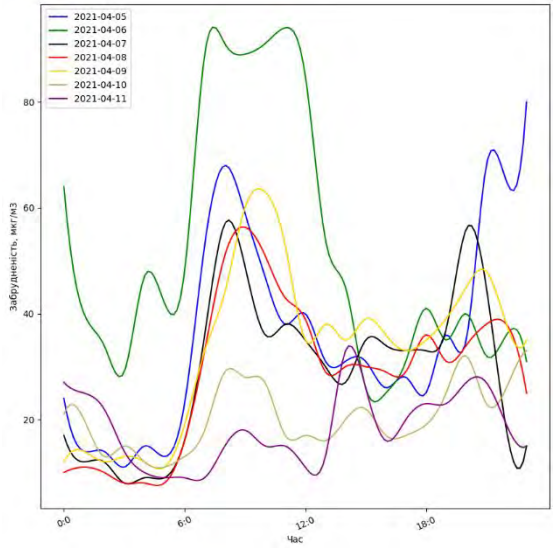


Рисунок 4.54 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 5.04-11.04.2021 на станції Escuelas Aguirre

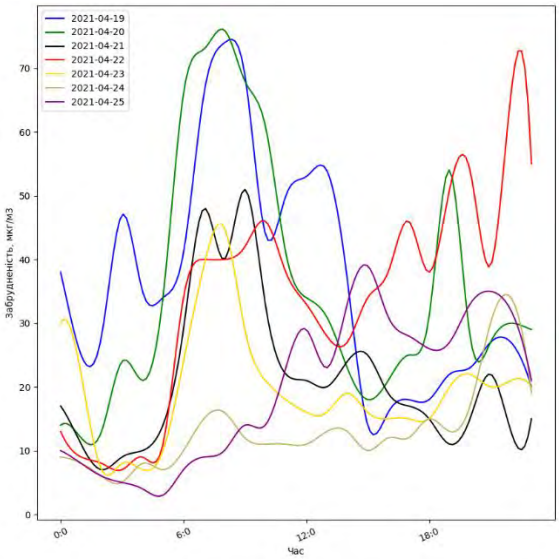


Рисунок 4.55 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 19.04-25.04.2021 на станції Vallecas

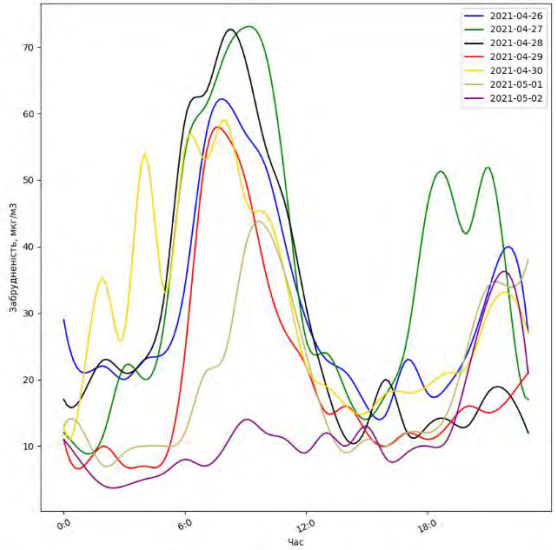


Рисунок 4.56 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 26.04-2.05.2021 на станції Vallecas

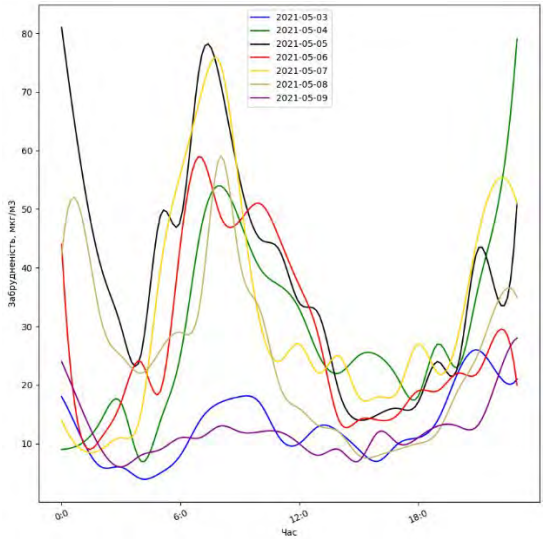


Рисунок 4.57 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 3.05-9.05.2021 на станції Vallecas

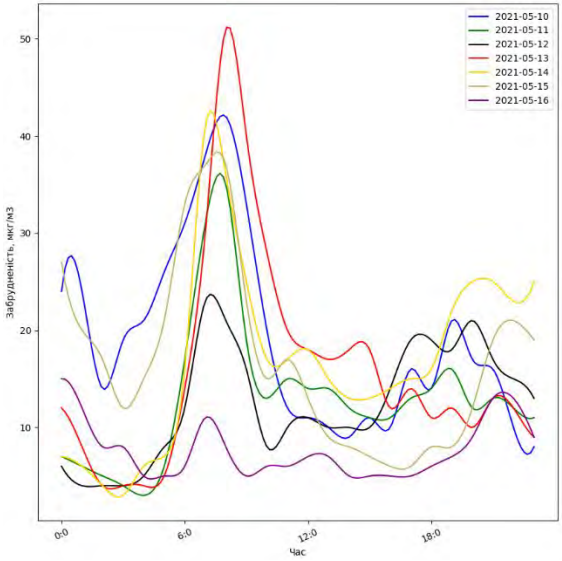


Рисунок 4.58 – Щогодинний рівень забрудненості атмосферного повітря діоксидом азоту 10.05-16.05.2021 на станції Vallecas

На рис. 4.59-4.66 продемонстровано на прикладі станції Avenida de Ramon y Cajal зміну рівня забрудненості в різні дати для днів тижня.

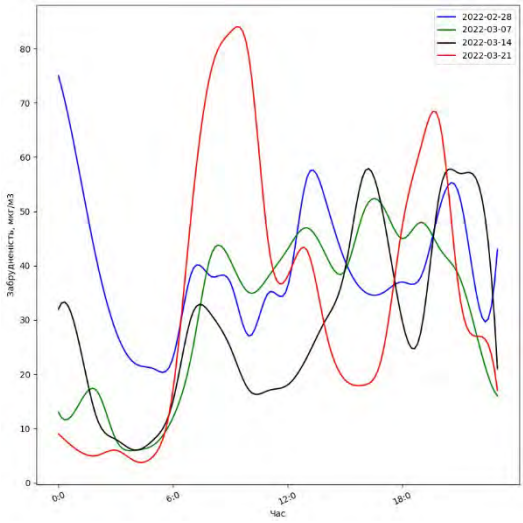


Рисунок 4.59 – Щогодинний рівень забрудненості повітря діоксидом азоту по понеділках на станції Avenida de Ramon y Cajal

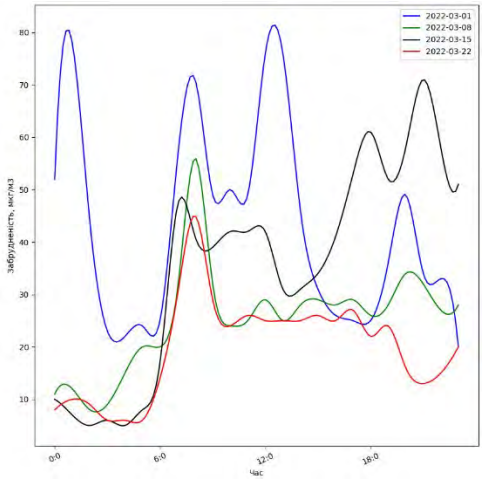


Рисунок 4.60 – Щогодинний рівень забрудненості повітря діоксидом азоту по вівторках на станції Avenida de Ramon y Cajal

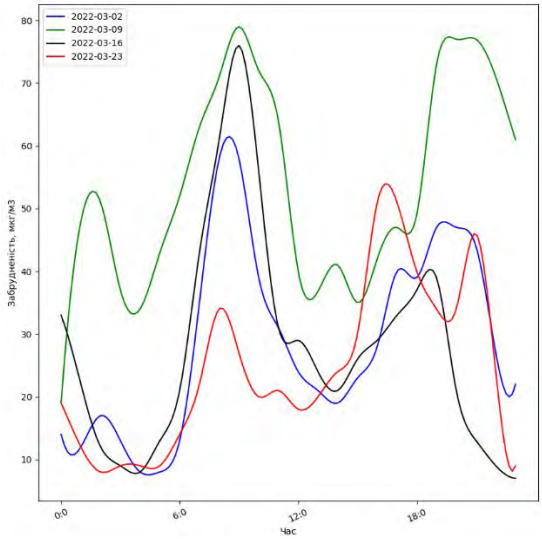


Рисунок 4.61 – Щогодинний рівень забрудненості повітря діоксидом азоту по середах на станції Avenida de Ramon y Cajal

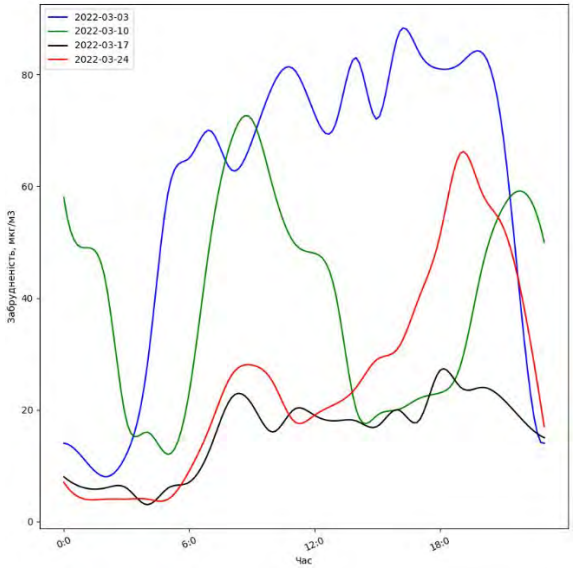


Рисунок 4.62 – Щогодинний рівень забрудненості повітря діоксидом азоту по четвергах на станції Avenida de Ramon y Cajal

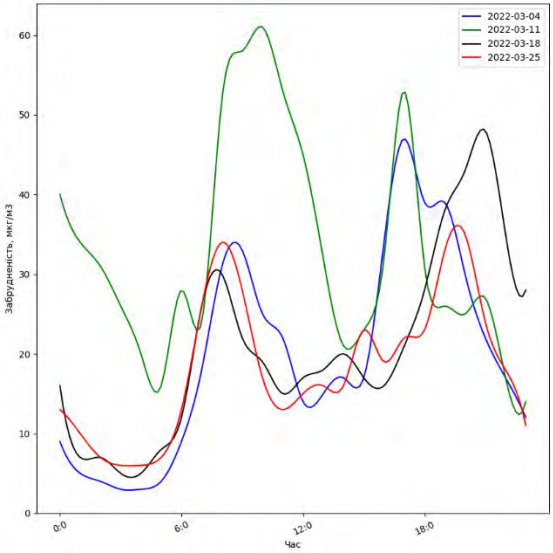


Рисунок 4.63 – Щогодинний рівень забрудненості повітря діоксидом азоту по п’ятницях на станції Avenida de Ramon y Cajal

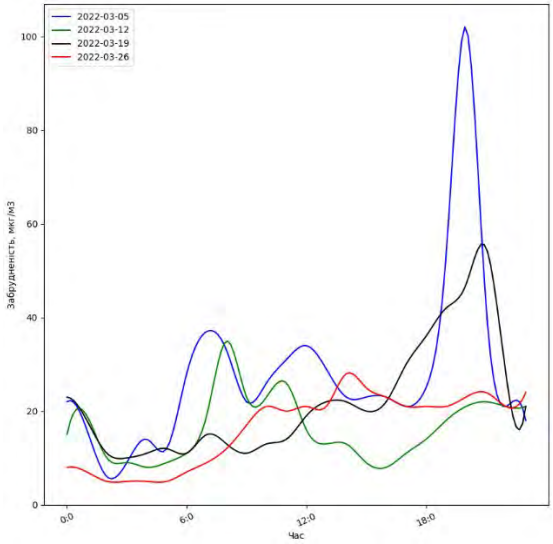


Рисунок 4.64 – Щогодинний рівень забрудненості повітря діоксидом азоту по суботах на станції Avenida de Ramon y Cajal

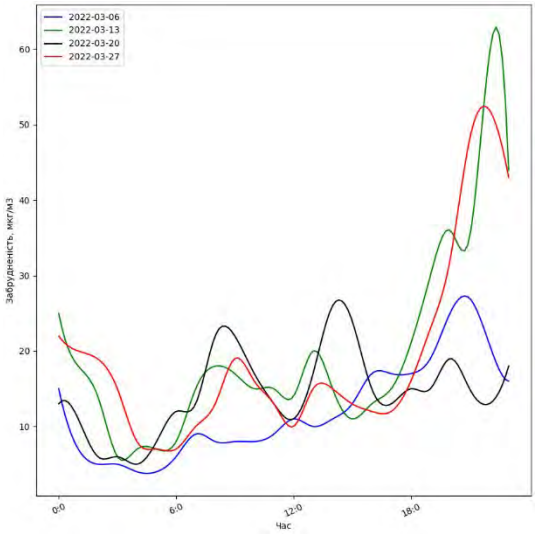


Рисунок 4.65 – Щогодинний рівень забрудненості повітря діоксидом азоту по неділях (6-27.03.2022) на станції Avenida de Ramon y Cajal

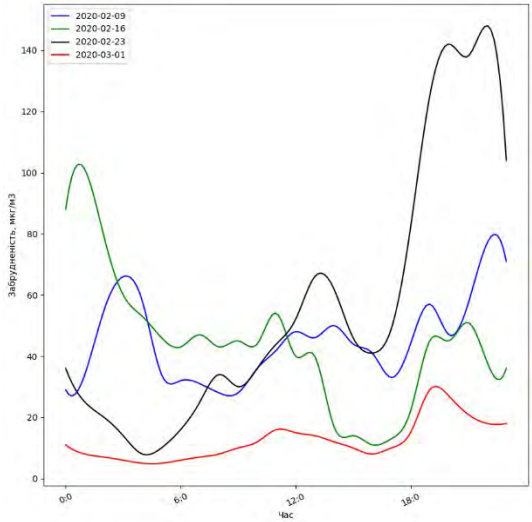


Рисунок 4.66 – Щогодинний рівень забрудненості повітря діоксидом азоту по неділях (9.02-3.01.2020) на станції Avenida de Ramon y Cajal

Приклад на рис. 4.66 цікавий тим, що на ньому значно виділяється

одна з наведених кривих. Крива, позначена червоним кольором, стосується 1 березня 2020 року. Цей момент фактично можна вважати певною невизначеністю перед офіційними карантинними обмеженнями під час пандемії COVID-19 у Іспанії. Дані, наведені на графіку, стосуються неділі, тобто демонструють те, як самі люди відреагували на таку невизначеність, адже з попереднього аналізу відомо, що рівень забрудненості повітря діоксидом азоту в неділю є щонайменше часто нижчим у вихідні дні, проте в даному випадку при загальному зниженні активності відбулось значне зниження саме в конкретну дату. Ця тенденція спостерігалась і на інших станціях, хоча на деяких з них не у кожну годину значення 1 березня 2020 року могло бути найнижчим.

Загалом проведений аналіз вказує на те, що на забрудненість повітря діоксидом азоту в Мадриді має вплив багато різних факторів. Визначити зміни тільки одним або певною невеликою обмеженою групою факторів буде достатньо некоректним, адже форма представлених функціональних залежностей значно відрізняється, хоча і є певні патерни поведінки, які можуть при тому не повторюватися між різними станціями, бути відмінними на одних і тих самих станціях у різний час. Тому в подальшому потрібно дослідити рівень впливу різних факторів окремо, що має стати передосновою побудови моделі прогнозування рівня забрудненості повітря діоксидом азоту.

Для проведення експериментального дослідження спочатку було сформовано відповідні вибірки даних. Формування вибірок було виконано на основі 2 окремих етапів:

– формування часових рядів за даними концентрації забруднювачів у атмосферному повітрі та метеорологічними даними за всіма станціями, на яких відбувається вимірювання концентрації забруднювачів;

– формування послідовності вхідних значень для кожного екземпляру, визначеного для першої вибірки, що включають автомобільний трафік за кожною стацією вимірювання трафіку за минулий період та за майбутній.

Набір метеорологічних даних було розширено за рахунок введення додаткового параметру на основі категоризації напрямку вітру. Для цього напрямок вітру був зведений до 8 можливих варіантів замість звичайного кута напрямку вітру, який був одним з інших

метеорологічних параметрів, які входили у вибірку.

На першому етапі було сформовано у підсумку вибірку, кожен ключ якої відповідає станції вимірювання концентрації забруднювача. За кожним ключем було створено окремі ключі, які відповідають номеру конкретного забруднювача та окремо метеорологічним даним. Метеорологічні дані були накопичені за кожною станцією окремо. Для цього для кожної станції вимірювання концентрації забруднювача за кожним метеорологічним параметром було визначено найближчу до неї метеорологічну станцію, яка здійснювала вимірювання цього параметра. Після цього відповідні результати були оброблені шляхом перевірки пропущених значень та заміни цих значень на основі лінійної інтерполяції. Далі ці дані були перетворені на основі нормування значень, для чого спочатку було визначено мінімальні та максимальні значення за кожним з параметрів, а тоді сформовано відповідний часовий ряд за кожним екземпляром, що визначає відповідний момент часу. За кожним таким екземпляром було визначено значення відповідного показника за 6 минулих годин.

Так само для кожної станції було виконано оброблення даних доступних за нею забруднювачів з формуванням на виході часового ряду після заповнення пропущених значень, нормування та формування ряду на основі значень за 6 минулих годин.

На другому етапі було сформовано окрему вибірку, кожен ключ якої містить номер станції вимірювання трафіку. За кожною станцією було за окремим ключем сформовано екземпляри на основі часового ряду, які відповідають даним за минулі 6 годин, та за окремим ключем дані, які відповідають наступним 6 годинам, які були обчислені за допомогою моделей прогнозування, які були сформовані за результатами попереднього підрозділу.

Дані було розбито на навчальну та тестову вибірки за тим же самим принципом, як це було зроблено в попередньому підрозділі.

Спершу було сформовано стандартні LSTM-моделі, які на основі значень забруднювача за минулі 6 годин прогнозують його значення на наступні 6 годин, використовуючи дані тільки однієї станції, для якої модель і будується. Відповідно таким чином було створено в підсумку 24 моделі. Отримані результати було оцінено та представлено в таблиці 4.7. Побудовані таким чином LSTM-моделі складаються з 3 прихованих шарів, зважаючи на те, що саме така архітектура дозволила отримати найкращі результати в підсумку,

використовуючи при цьому між шарами виключення (dropout) на рівні 0,5 та 0,4 відповідно.

Таблиця 4.7 – Результати прогнозування забрудненості атмосферного повітря діоксидом азоту на основі базових LSTM-моделей

Станція	MSE	MAE	RMSE
1	2	3	4
Plaza de Espana	0,00247	0,03156	0,04971
Escuelas Aguirre	0,00263	0,03585	0,05129
Ramon y Cajal	0,00489	0,04641	0,06995
Arturo Soria	0,00338	0,03634	0,05818
Villaverde	0,0085	0,06193	0,09217
Farolillo	0,00526	0,05002	0,07251
Casa de Campo	0,00464	0,04449	0,06809
Barajas Pueblo	0,00571	0,05408	0,07557
Plaza del Carmen	0,00561	0,05137	0,07489
Moratalaz	0,00484	0,04794	0,06959
Cuatro Caminos	0,00504	0,04491	0,07098
Barrio del Pilar	0,00709	0,05863	0,0842
Vallecas	0,00608	0,05451	0,078
Mendez Alvaro	0,00812	0,06211	0,09011
Castellana	0,00621	0,05283	0,07883
Parque del Retiro	0,00504	0,0454	0,071
Plaza Castilla	0,00542	0,05254	0,07364
Ensanche de Vallecas	0,0025	0,03527	0,05003
Urb. Embajada	0,0046	0,04348	0,06784
Plaza Elíptica	0,00412	0,04331	0,06417
Sanchinarro	0,00751	0,05669	0,08664

Продовження таблиці 4.7

1	2	3	4
El Pardo	0,00371	0,04317	0,0609
Juan Carlos I	0,00554	0,05001	0,0744
Tres Olivos	0,00541	0,05213	0,07358

Враховуючи показники, які використовуються в задачі (4.2) і визначають відповідну модель, необхідно було визначити спосіб формування підмножин K^S, R^S, B^C. Для виконання такого відбору було використано підхід, який ґрунтується на застосуванні ансамблів дерев рішень за допомогою методу Random Forest.

Однак перед застосуванням даного методу спочатку було проаналізовано наявний зв'язок між значеннями різних забруднювачів. Для цього було об'єднано дані за всіма станціями. Для аналізу використовувалися дані, починаючи з 2001 року, не включаючи дані, що входили до тестової вибірки. На рис. 4.67 показано результати обчислення коефіцієнту Пірсона на основі даних за поточні години. Тобто взаємозалежність встановлювалася між значеннями різних показників, але за одну й ту саму поточну годину.

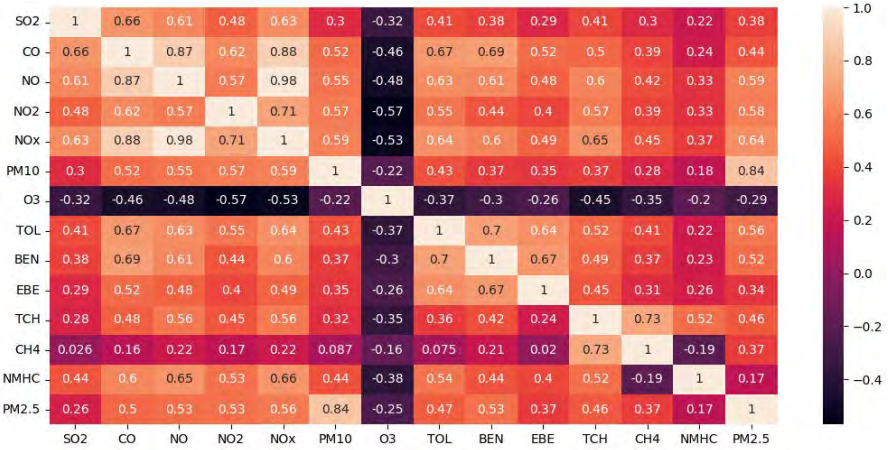


Рисунок 4.67 – Взаємна кореляція між значеннями забруднювачів за всіма станціями за поточну годину

На рис. 4.68 показана відповідні розрахунки при зміщенні на годину. Тобто взаємозалежність встановлюється між значеннями забруднювача за поточну годину та значеннями іншого забруднювача за попередню годину. Одразу видно зменшення значень відповідних коефіцієнтів, хоча при цьому загальна тенденція залишається достатньо вираженою.

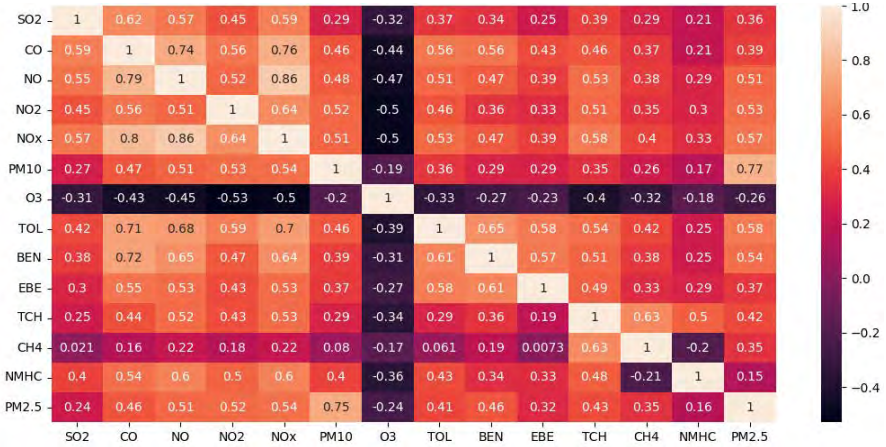


Рисунок 4.68 – Взаємна кореляція між значеннями забруднювачів за всіма станціями за попередню годину

Детальний аналіз виконувався окремо за станціями, що дозволило зокрема зрозуміти, що значення відповідних коефіцієнтів кореляції можуть значно відрізнятися, а до того ж додатково продемонструвати різницю в кількості забруднювачів, дані за якими доступні за різними станціями. Наприклад, на рис. 4.69 показані відповідні розрахунки за станцією Plaza de Espana, а на рис. 4.70 – за станцією Escuelas Aguirre. Множина доступних забруднювачів за цими станціями значно відрізняється. Окрім того і значення відповідних коефіцієнтів за діоксидом азоту за станцією Plaza de Espana значно менші ніж за станцією Escuelas Aguirre, де відповідні залежності від інших забруднювачів є більш вираженими.

Відповідно є необхідність вибору забруднювачів у розрізі конкретних станцій та конкретних станцій, на яких концентрація цих забруднювачів вимірювалась для заданої станції прогнозування.

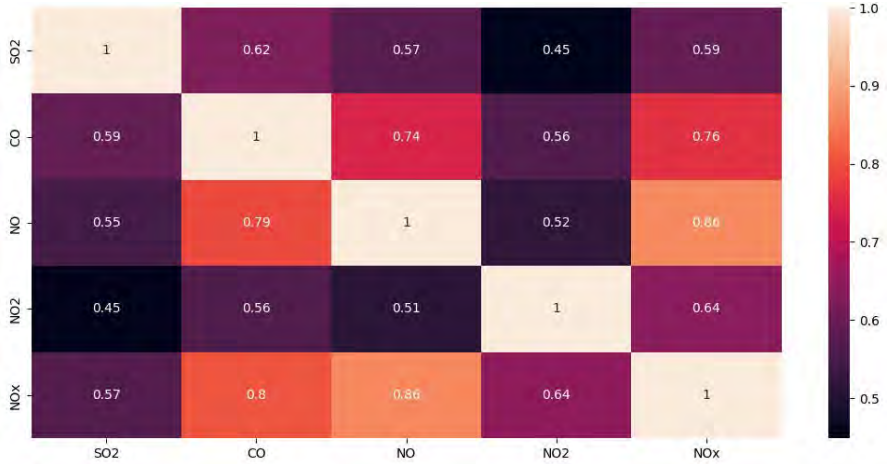


Рисунок 4.69 – Взаємна кореляція між значеннями забруднювачів за станцією Plaza de Espana за попередню годину

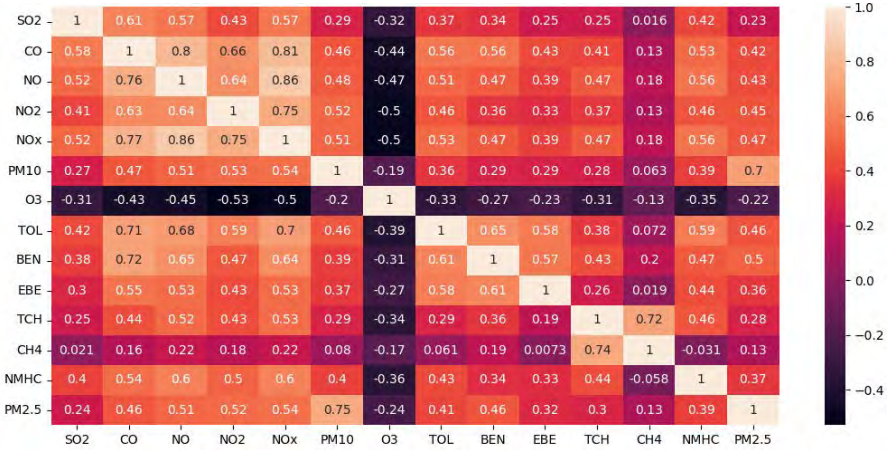


Рисунок 4.70 – Взаємна кореляція між значеннями забруднювачів за станцією Escuelas Aguirre за попередню годину

Окрім того було проаналізовано автокореляцію даних діоксиду азоту (рис. 4.71-4.72).

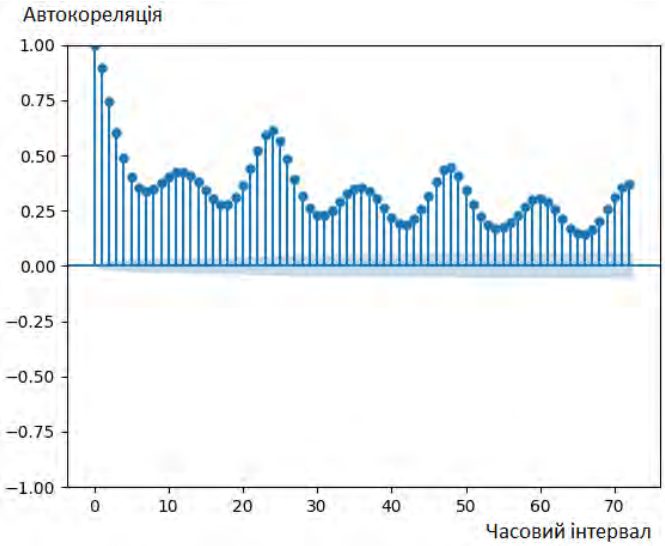


Рисунок 4.71 – Графік автокореляції концентрації діоксиду азоту за станцією Plaza de Espana

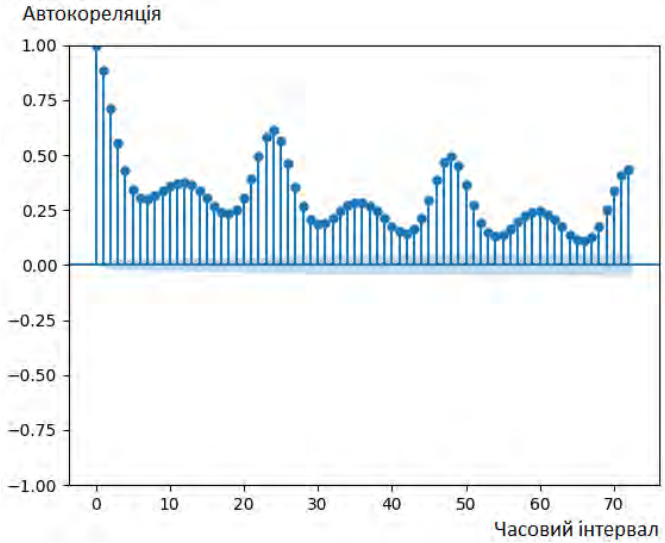


Рисунок 4.72 – Графік автокореляції концентрації діоксиду азоту за станцією Barajas Pueblo

На рис. 4.71 та 4.72 можна помітити для різних станцій спільну тенденцію. При цьому помітним також є те, що підвищення автокореляції відбувається ближче до кінця доби, але при цьому пікове значення є відносно невисоким. Якщо порівнювати з результатами для попередньої задачі (4.1), то значення є значно нижчими. Відповідно фактично загальна тенденція формується протягом 6 годин, а далі не досягає вже таких піків. За жодною станцією автокореляція за 24 години не досягає 0,75. Тому можна зробити висновок, що зі збільшенням довжини часового ряду ймовірно можна отримати додаткову інформацію, проте вона має значно меншу інформативність при прогнозуванні. Тому для випадку зменшення складності моделей достатньо побудувати модель на основі 6 попередніх годин.

Другим типом моделей, які були створені, були LSTM-моделі, які отримували на вхід не тільки дані концентрації діоксиду азоту за станцією прогнозування за минулий період, але і дані про вмісті інших забруднювачів за іншими станціями за той же самий період. Виходячи з результатів аналізу взаємозалежностей, ґрунтуючись на експериментальних дослідженнях, були виділені дані про концентрацію діоксиду азоту, але на інших станціях, та дані про концентрацію оксидів азоту. Для визначення конкретних станцій, за якими подавалися значення цих забруднювачів, було використано ансамблі дерев рішень за допомогою методу Random Forest. Для кожної станції за кожним з цих 2 забруднювачів виділялись максимально 3 станції, за якими могли подаватися дані на вхід. У підсумку даного застосування за кожною станцією було визначено відповідні станції для кожного з цих 2 забруднювачів. Максимально відповідно 6 додаткових ознак могла модель отримати на вхід. Для деяких станцій було виділено в підсумку дані тільки за одним забруднювачем. На основі цієї виділеної множини вхідних ознак для кожної станції було побудовано LSTM-моделі, які при цьому вже мали тільки 2 приховані шари.

Третім типом моделей були моделі на основі LSTM, які окрім значення забрудненості повітря діоксидом азоту за станцією прогнозування, отримували також дані автомобільного трафіку за минулий період у 6 годин. Для цього також було використано метод Random Forest, який дозволив виділити станції вимірювання трафіку, які є релевантними для кожної такої станції прогнозування.

Четвертий тип моделей був аналогічним попередньому, але на вхід додатково подавались дані трафіку не за минулий період, а майбутні прогнозовані значення на основі моделей, сформованих у попередньому підрозділі. Для них також за кожною станцією було обрано перелік станцій вимірювання трафіку, дані яких є релевантними. При таких обчисленнях використовувались дані прогнозування, застосовані тільки для навчальної вибірки.

П’ятий тип моделей мав у якості додаткових ознак до однієї ознаки базової LSTM-моделі метеорологічні показники. На вхід подавалися всі показники окрім напрямку вітру, замінюючи його категоріальним значенням.

Шостий тип моделей був результуючим і об’єднував на вхід LSTM-моделей прогнозовані значення трафіку за рядом станцій, минулі значення концентрації діоксиду азоту та оксидів азоту за рядом станцій. Отримані результати зведені до таблиці 4.8.

Таблиця 4.8 – Результати прогнозування забрудненості атмосферного повітря діоксидом азоту на основі результуючих LSTM-моделей

Станція	MSE	MAE	RMSE
1	2	3	4
Plaza de Espana	0,00278	0,03642	0,05269
Escuelas Aguirre	0,00226	0,03213	0,04756
Ramon y Cajal	0,00421	0,04231	0,06491
Arturo Soria	0,00357	0,03922	0,05972
Villaverde	0,0075	0,05863	0,08662
Farolillo	0,00475	0,04687	0,06893
Casa de Campo	0,00373	0,0401	0,06109
Barajas Pueblo	0,005	0,04982	0,07074
Plaza del Carmen	0,00502	0,04849	0,07085
Moratalaz	0,00422	0,04398	0,06495
Cuatro Caminos	0,00484	0,04284	0,06954
Barrio del Pilar	0,00608	0,05361	0,07798
Vallecas	0,00554	0,0512	0,07445

Продовження таблиці 4.8

1	2	3	4
Mendez Alvaro	0,00733	0,05987	0,08561
Castellana	0,00554	0,04955	0,0744
Parque del Retiro	0,00468	0,04249	0,06843
Plaza Castilla	0,00486	0,04834	0,0697
Ensanche de Vallecas	0,00241	0,03382	0,04912
Urb. Embajada	0,00422	0,04058	0,06499
Plaza Elíptica	0,00383	0,04065	0,06186
Sanchinarro	0,00654	0,05385	0,08085
El Pardo	0,00321	0,03886	0,05666
Juan Carlos I	0,00486	0,04711	0,06973
Tres Olivos	0,00477	0,04847	0,06907

Усі підсумкові середні результати, отримані за всіма 24 моделями, побудованими за описаними вище принципами, представлені в таблиці 4.9.

Таблиця 4.9 – Середні результати прогнозування рівня забрудненості атмосферного повітря діоксидом азоту на наступні 6 годин на основі різних моделей

Модель прогнозування	MSE	MAE	RMSE
LSTM	0,00518	0,04812	0,07109
LSTM з додатковими забруднювачами	0,00499	0,04842	0,06994
LSTM з трафіком за минулий період	0,00511	0,04668	0,07068
LSTM з прогнозованим трафіком	0,00485	0,04571	0,0689
LSTM з метеорологічними даними	0,00565	0,05167	0,07434
Запропонована модель	0,00466	0,04538	0,06752

Отримані результати демонструють, що в результаті використання додаткових ознак на основі забруднювачів з релевантних станцій вдалось дещо покращити результати базової LSTM-моделі з 1

вхідною ознакою. Також невелике покращення результатів було отримано за використання даних з релевантних станцій про рівень автомобільного трафіку. При цьому за такого використання середня абсолютна похибка збільшилась на 3 %, а MSE лише на 1,35 %. Тож в даному випадку результати вказують на наявність певної додаткової інформації в даних трафіку, але дані за минулий період є достатньо застарілими для прогнозування. У свою чергу метеорологічні дані за використання всього набору призвели до погіршення результатів (мають багато зайвого шуму), потребують вивчення в подальшому того факту, чи є серед них окремі показники, які можуть бути релевантними.

LSTM-моделі, побудовані на використанні результатів прогнозування трафіку за релевантними станціями, дозволили на 6,37 % зменшити значення похибки MSE порівняно з базовими LSTM-моделями. Використання ж разом даних про концентрацію релевантних забруднювачів за релевантними станціями разом з прогнозованими значеннями трафіку в підсумкових моделях призвело до зменшення середньоквадратичної похибки MSE на 10,04 %. Тож об'єднання моделей прогнозування у фреймворку прийняття рішень для медичного діагностування дозволило отримати більш надійну інформацію для прийняття рішень на останньому етапі в підсумку, що є важливим для практичного застосування.

4.4 Висновки за розділом 4

У даному підрозділі було запропоновано фреймворк прийняття рішень для медичного діагностування, що дозволяє об'єднати створені моделі в єдину систему, визначити взаємодію між ними та створити єдине сховище даних. Проведено експериментальне дослідження, яке в підсумку охоплювало задачі прогнозування автомобільного трафіку та прогнозування забрудненості атмосферного повітря. Об'єднання цих моделей у складі фреймворку дозволило збільшити в підсумку точність прогнозування забрудненості атмосферного повітря діоксидом азоту. Створені в результаті дослідження моделі дозволяють виконувати прогнозування в умовах відсутності великої системи станцій як для вимірювання автомобільного трафіку, так і для вимірювання концентрації забруднювачів у повітрі, а також обмеженості накопичених історичних даних.

4.5 Література до розділу 4

1. Urban traffic flow prediction techniques : A review / Boris Medina-Salgado, Eddy Sánchez-DelaCruz, Pilar Pozos-Parra, Javier E. Sierra // Sustainable Computing : Informatics and Systems. – 2022. – Volume 35. – 100739. – DOI : 10.1016/j.suscom.2022.100739.

2. Traffic prediction based on GCN-LSTM model / Zhizhu Wu, Mingxia Huang, Aiping Zhao and Zhixun lan // Journal of Physics: Conference Series. – 2021. – Volume 1972. – 012107. – DOI : 10.1088/1742-6596/1972/1/012107.

3. Chen, R. Hybrid Graph Models for Traffic Prediction / R. Chen, H. Yao // Applied Sciences. – 2023. – № 13 (15). – 8673. – DOI : 10.3390/app13158673.

4. Yu, B. Spatio-Temporal Graph Convolutional Networks : A Deep Learning Framework for Traffic Forecasting / Bing Yu, Haoteng Yin, Zhanxing Zhu // Proceedings of the 27th International Joint Conference on Artificial Intelligence. – 2018. – Pp. – 3634-3640. – DOI : 10.48550/arXiv.1709.04875.

5. Attention-based Conv-LSTM and Bi-LSTM networks for large-scale traffic speed prediction / X. Hu, T. Liu, X. Hao et al. // The Journal of Supercomputing. – 2022. – Volume 78. – Pp. 12686–12709. – DOI : https://doi.org/10.1007/s11227-022-04386-7.

6. Fu, R. Using LSTM and GRU neural network methods for traffic flow prediction / R. Fu, Z. Zhang and L. Li // 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China. – 2016. – Pp. 324-328. – DOI : 10.1109/YAC.2016.7804912.

7. Abduljabbar, R.L. Development and evaluation of bidirectional LSTM freeway traffic forecasting models using simulation data / R.L. Abduljabbar, H. Dia, P.-W. Tsai. – Scientific Reports. – 2021. – Volume 11. – 23899 – DOI : 10.1038/s41598-021-03282-z.

8. LSTM network : a deep learning approach for short-term traffic forecast / Zheng Zhao, Weihai Chen, Xingming Wu, Peter C. Y. Chen, Jingmeng Liu // IET Intelligent Transport Systems. – 2017. – Volume 11, Issue 2. – Pp. 68-75.

9. Poonia, P. Short-Term Traffic Flow Prediction : Using LSTM / Pregya Poonia, V. Jain // 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3). – 2000. – 1-4. – DOI : 10.1109/ICONC345789.2020.9117329.

10. En portada – Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/portal/site/egob.

11. Aforos de tráfico en la ciudad de Madrid permanentes - Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/sites/v/index.jsp?vgnextoid=fabbf3e1de124610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

12. Datos meteorológicos. Datos horarios desde 2019 - Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=fa8357cec5efa610VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default.

13. Lovkin, V. Air Pollution Prediction as a Source for Decision Making Framework in Medical Diagnosis / V. Lovkin, A. Oliinyk, Y. Lukashenko // IntelITSIS'2021: 2nd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 24–26, 2021, Khmelnytskyi, Ukraine. – Khmelnytskyi : KhNU, 2021. – Pp. 295-302.

14. Information Model of Outdoor Air Pollution Prediction for Medical Diagnosis System / V. Lovkin, A. Oliinyk, T. Fedoronchak, Y. Lukashenko // 4th IEEE International Conference on Advanced Information and Communication Technologies (AICT) - 2021, September 21–25, 2021, Lviv, Ukraine. – Lviv : LPNU, 2021. – Pp. 141-144.

15. PM2.5 Air Pollution Prediction through Deep Learning Using Multisource Meteorological, Wildfire, and Heat Data / Pratyush Muthukumar, Kabir Nagrecha, Dawn Comer et al. // Atmosphere. – 2022. – Volume 13, Issue 5. – 822. – DOI : 10.3390/atmos13050822.

16. Kaya, K. Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting / K. Kaya, Sule Gundüz Oguducu // Scientific Reports. – 2020. – Volume 10. – 3346. – DOI : 10.1038/s41598-020-60102-6.

17. AlShehhi, A. Artificial intelligence for improving Nitrogen Dioxide forecasting of Abu Dhabi environment agency ground-based stations / A. AlShehhi // Journal of Big Data. – 2023. – Volume 10. – 92.

18. Xayasouk, T. Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models / Thanongsak Xayasouk, HwaMin Lee, Giyeol Lee // Sustainability. – 2020. – 12 (6). – 2570. – DOI : 10.3390/su12062570.

19. Iskandaryan, D. Bidirectional convolutional LSTM for the prediction of nitrogen dioxide in the city of Madrid / Ditsuhi Iskandaryan, Francisco Ramos, Sergio Trilles // PLoS One. – 2022. – Volume 17, Issue 6. – e0269295. – DOI : 10.1371/journal.pone.0269295.

20. Calidad del aire. Datos horarios desde 2001 - Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=f3c0f7d512273410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default.

21. Calidad del aire. Datos diarios desde 2001 - Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/sites/v/index.jsp?vgnextoid=aecb88a7e2b73410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

22. Calidad del aire. Estaciones de control - Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=9e42c176313eb410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default.

23. Datos meteorológicos. Estaciones de control - Portal de datos abiertos del Ayuntamiento de Madrid [Електронний ресурс]. – Режим доступу : https://datos.madrid.es/sites/v/index.jsp?vgnextoid=2ac5be53b4d2b610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD.

24. WHO global air quality guidelines : particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide [Електронний ресурс]. – 290 p. – Режим доступу : https://www.who.int/publications/i/item/9789240034228.

Наукове видання

Субботін Сергій Олександрович
Олійник Андрій Олександрович
Льовкін Валерій Миколайович
Леощенко Сергій Дмитрович

Інтелектуальні методи, фреймворки та програмні засоби для прогнозування і діагностування нелінійних об'єктів

Монографія

Комп'ютерний набір *Субботін С.О.*
Верстання *Субботін С.О.*

Підписано до друку 07.12.2023. Формат 60×84/16. Ум. друк. арк. 13,19.
Тираж 100 прим. Зам. № 1008.

Національний університет «Запорізька політехніка»
Україна, 69063, м. Запоріжжя, вул. Жуковського, 64
Тел.: (061) 769–82–96, 220–12–14

Свідоцтво суб'єкта видавничої справи ДК № 6952 від 22.10.2019.